

半教師あり学習を用いた植物生育モデリングの検討

水野涼介¹ 柴田瞬¹ 峰野博史²

概要: 国内の農業分野において、熟練農家の高度な栽培技術の喪失の課題を解決するために、経験と勘に頼っていた栽培技術を形式知化する研究が進められている。高度な栽培技術の中でも、植物の状態に応じて、適切なストレスを付与するストレス栽培では高糖度な果実を生産できることが知られており、このストレス栽培の形式知化を目的とした様々な植物水分ストレス推定手法が提案されてきた。水分ストレスの推定結果に基づき灌水制御を自動化することで熟練農家の負担を軽減するだけでなく、新規就農者でも高品質な果実の安定生産が期待されている。一方で、一般に熟練農家の高度な果実栽培では、植物の生育状態に基づいて収穫時の果実品質を調節していると言われていたが、果実糖度などの植物生育状態の指標は膨大なラベル付きデータの収集が困難であり、膨大なラベル付きデータを必要とする既存手法では収穫時の品質調節を目的とした植物生育モデリングは困難であった。そこで、本研究では少量のラベル付き植物生育状態データからでも、半教師あり学習を用いて高精度に植物の生育状態をモデル化できる手法を検討する。トマトの積算蒸発散量を目的変数とした基礎評価の結果、提案手法はMLPに比べMAEで約37.9%、Stacked LSTMに比べ約25.6%の誤差を削減できるだけでなく、膨大なラベル付きデータが存在しないことを想定し、ラベル付きデータを最大75%削減した場合でも、推定精度を維持できることを確認した。

Study on plant growth modeling using semi-supervised learning

RYOSUKE MIZUNO¹ SHUN SHIBATA¹ HIROSHI MINENO²

1. はじめに

近年、国内の農業従事者人口や若年層割合が低下しており、熟練農家の高度な栽培技術の喪失が課題となっている[1]。国内の熟練農家の高度な栽培技術は世界的に高水準な栽培技術であり、高度な栽培技術は、熟練農家が長い年月をかけて培った経験と勘に基づいて行われている。そのため、栽培技術継承のためには経験と勘を十分に培う必要があり、経験と勘が十分に培われていない新規就農者への継承には時間を要する。この課題を解決するために、ICT (Information and communication technology) を用いて、高度な栽培技術を形式知化する研究が進められている。栽培技術を形式知化することで、新規就農者であっても、高度な栽培技術の知見を得ることが可能である。さらには、自動化することで農業従事者人口の多くを占めている、高齢者の負担の軽減、新規就農者であっても高品質な果実の安定生産や安定した収入を得ることが可能である。高品質な果実や収入を安定して得ることが実現することで、若年層による農業のマイナスイメージとして多くあげられる、農業は重労働である点、所得が不安定である点[2]が解消され、新規就農者の増加が期待できる。

高度な栽培技術の中でも、植物の状態に応じて適切なストレスを付与するストレス栽培では、高糖度トマトに代表

される高糖度な果実を生産できることが知られており、果実の高品質化、果実の市場価値の向上につながることから、形式知化が期待されている。このストレス栽培の形式知化を目的とした様々な植物水分ストレス推定手法が提案されてきた。これまでに、植物の茎径や重量から水分ストレスを推定する手法[3]や温度、湿度などの環境データや草姿画像データ、機械学習を用いて水分ストレスを推定し、推定結果に基づき灌水制御を自動化する手法が提案されている[4, 5]。一方で、一般に熟練農家の行っている高度な果実栽培は、植物生育状態に基づいて収穫時の糖度を調節していると言われていた。また、ICTの発展した現在、温度や湿度などの環境データや茎径、重量などの生育データの自動計測技術は確立されつつあるが、経時的な糖度の自動計測技術は現状では確立されていない。そのため、植物生育状態の指標である果実糖度は膨大なラベル付きデータの収集が困難であり、膨大なラベル付きデータを必要とする既存手法では収穫時の品質調節を目的とした植物生育モデリングが困難であった。

そこで本研究では、半教師あり学習を用いた植物生育モデリング手法を検討する。これまでに、半教師あり学習を用いた植物生育モデリング手法として、Semi-supervised deep state-space model (SDSSM) [6]が提案されている。SDSSMは少量の教師データの下でも、植物状態や気象環境と経時糖度間の複雑な関係性をモデル化し、熟練農家による収穫時果実品質の調節を適切に再現する。また、SDSSMはNeural network (NN) を用いて状態空間モデル (SSM: State

1 静岡大学大学院総合科学技術研究科
Graduate School of Integrated Science and Technology, Shizuoka University
2 静岡大学学術院領域/JST さきがけ
College of Informatics, Academic Institute, Shizuoka University/JST, PRESTO

space model) を表現したモデルであり、少量のラベル付きデータと膨大な量のラベルなしデータを用いてデータの複雑な生成過程や状態遷移をモデル化が可能である。ここで、SDSSM はラベル付きデータが少量であっても、適切に植物生育状態をモデル化できるとしたら、SDSSM は経時糖度を指標とした植物生育モデリングに有効であることが明らかになる。しかし、これまでに行われた SDSSM の性能評価は少量の経時糖度データと一定量の訓練データを用いた評価のみであった。そこで、本稿ではトマト栽培時の新たな実経時データを用いた積算蒸発散量推定を行い、既存手法との推定精度比較、訓練データ数に応じた SDSSM の推定性能評価を行うことで半教師あり学習を用いた植物生育モデリングの検討を行う。

以降、2 章では関連技術及び研究に関して述べ、3 章では半教師あり学習を用いた植物生育モデリング手法を述べる。4 章では本手法の基礎評価結果を述べ、最後に 5 章で本稿をまとめる。

2. 関連技術及び研究

植物生育モデリングの関連研究として、SSM を用いて、小麦の成長プロセスをモデル化する研究[7]がある。植物の生育データをはじめとして、時系列性を持つデータは多くのノイズを含み、高次元なデータであることが知られている[8]。一方、SSM はノイズに対して頑健、推定値の曖昧性が評価可能なモデルであり、状態モデルと観測モデルと呼ばれる 2 つのモデルで表される時系列統計モデルである。SSM では直接的に観測出来る観測値を観測変数、直接的に観測出来ない潜在的な状態を潜在変数と仮定することで、観測値と状態の 2 種類のモデル化を可能としている。ここで、観測変数を x_1, x_2, \dots, x_T 、潜在変数を z_1, z_2, \dots, z_T とすると、SSM のグラフィカルモデルは図 1 (a) のように表現され、生成過程は以下の式 (1)、式 (2) で示される。

$$z_t \sim p_\theta(z_t | z_{t-1}) \quad (1)$$

$$x_t \sim p_\varphi(x_t | z_t) \quad (2)$$

式 (1)、式 (2) はそれぞれ状態モデル、観測モデルである。時点 t ($t=0, 1, \dots, T$) において、潜在変数 z_t は 1 時点前の潜在変数 z_{t-1} の条件付き確率分布から生成され、観測変数 x_t は潜在変数 z_t から生成される。また、 θ 、 φ はそれぞれ状態モデル、観測モデルのパラメータであり、状態モデルにおける潜在変数の初期値 z_0 は $x_0 = p(x_0)$ で初期化が行われる。状態空間モデルにおいて最も単純な線形ガウス状態空間モデル[9]は以下の式 (3)、式 (4) で表される。

$$z_t = F_t z_{t-1} + v_t \quad (3)$$

$$x_t = H_t z_t + w_t \quad (4)$$

式 (3)、式 (4) における F_t 、 H_t はそれぞれモデルの表現形式を決める係数行列であり、 v_t 、 w_t はノイズである。ここで、植物状態を潜在状態、植物栽培環境の温度や湿度、

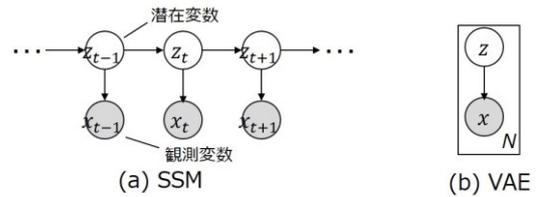


図 1 SSM, VAE のグラフィカルモデル

植物の茎径を観測値とすることで、本来観測できない状態である、植物状態のモデル化が期待できる。

また、状態空間モデルによる時系列解析とその生態学への応用に関する研究[10]において、深谷は、状態空間モデルの観測過程と背後にある状態過程・パラメータの構造を分離したモデル化によって、生態的過程に関する自然な推測を実現でき、状態空間モデルは生態学においてその重要性を増していくことを示した。しかし、一般的な状態空間モデルはデータ構造に線形性を仮定しており、複雑な植物のモデルをはじめとして実問題への応用には表現力が不十分であると言える。そこで、非線形性を持つデータ構造に対して時系列データをモデル化する手法として拡張カルマンフィルタ[11]や、非線形なデータ構造における複雑な生成過程を持つデータに対しモデル化する手法として変分自己符号化器 (VAE: Variational auto encoder) [12]が提案されている。VAE は非線形性を持つデータの深層生成モデルの 1 つであり、状態空間モデルと同様に観測変数 x と潜在変数 z を考えた時、生成過程は式 (5)、式 (6) で表され、グラフィカルモデルは図 1 (b) で表される。

$$z \sim p(z) \quad (5)$$

$$x \sim p_\theta(x|z) \quad (6)$$

式 (6) における θ はパラメータであり、潜在変数 z は $p(z)$ 、観測変数 x は $p_\theta(x|z)$ から生成される。この時、対数周辺尤度は式 (7) で表され、式 (7) を最大化するように学習される。

$$\log p_\theta(x) = \log \int p_\theta(x|z)p(z)dz \quad (7)$$

しかし、 $p_\theta(x)$ は解析的に計算困難であるため、 $p_\theta(x|z)$ の近似事後分布として、 $q_\varphi(z|x)$ を用いて式 (8) で示す変分下限 (ELBO: Evidence lower bound) を最大化することで学習を行う。

$$\begin{aligned} \log p_\theta(x) \\ \geq E_{q_\varphi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\varphi(z|x) || p(z)) \end{aligned} \quad (8)$$

ここで、式 (8) において D_{KL} は Kullback-Leibler divergence であり、事前分布 $p(z)$ と近似事後分布 $q_\varphi(z|x)$ の差異を示す尺度である。ここで、式 (8) における第 1 項は近似事後分布 $q_\varphi(z|x)$ からサンプリングを行った期待値であり、微分可能でないため、最適化時に勾配を求めることが出来ない。そのため、 $z \sim q_\varphi(z|x)$ を確率的なノイズ ϵ を用いて、 $\bar{z} = \mu(x) + \epsilon\sigma(x)$ と再パラメータ化することで、期待値はモン

テカルロサンプリングによって $E_{q_\phi(z|x)}[\log p_\theta(x|z)] \cong \frac{1}{L} \sum \log p_\theta(x|\bar{z})$ と近似でき、勾配を求めることが可能となる。

ここで、従来の VAE では困難であった時系列データを扱うため、STORN (Stochastic recurrent networks) [13] や DKF (Deep Kalman filters) [14] が提案されてきた。しかし、従来の手法では、植物データをはじめとして、実世界における複雑なデータ生成過程のモデル化には膨大な量のデータが必要であった。そのため、データ収集コストの高い農業分野においては、大量のデータ収集が困難であり、少量のラベルデータで高い表現力を持ったモデル化手法が必要となる。

3. 提案手法

3.1 概要

本研究では、膨大な量のデータ収集が困難な経時糖度を用いた植物生育モデリングを実現するため、半教師あり学習モデル化手法である SDSSM [6] を用いた植物生育モデリング手法の検討を行う。SDSSM は NN を用いて SSM を表現したモデルであり、少量のラベル付きデータと膨大な量のラベルなしデータを用いてデータの複雑な生成過程や状態遷移をモデル化が可能である。SDSSM のグラフィカルモデルを図 2 に示す。図 2 において z_t は潜在変数であり、 x_t はラベルデータとしての観測変数、 y_t は x_t と潜在変数 z_t を共有する観測変数、 u_t 、 s_t 、 r_t はそれぞれ z_t 、 x_t 、 y_t に作用する観測変数である。ラベルデータ y_t と観測変数 x_t は潜在変数 z_t を共有していることから、ラベルデータ x_t が欠損していても、潜在変数を共有する別の観測変数 y_t で潜在変数 z_t を更新することを可能とする。また、SDSSM は観測変数 x_t 、 y_t と潜在変数 z_t に対して、それぞれ異なる外部要因が設定されていることから、モデル化時に各変数の生成過程に適したデータを設定することで、適切に植物生育モデリングを行うことが可能であると考える。

そのため、SDSSM を用いたモデル化時に、潜在変数 z_t に植物状態、ラベルデータ x_t に糖度、ラベルデータと潜在変数を共有する観測変数 y_t に糖度と類似した植物状態を表すデータ、その他の外部変数 u_t 、 s_t 、 r_t に環境データや生育データを適切に設定することで、糖度データが少量であっても、植物生育モデルを構築することが可能であると考える。

3.2 SDSSM の目的関数と定式化

本節では、SDSSM の目的関数と定式化について述べる。図 2 における SDSSM の生成過程を式 (9)、式 (10)、式 (11) に示す。

$$p_\theta(z_t|z_{t-1}, u_t) = N(z; \mu_z(z_{t-1}, u_t), \sigma_z(z_{t-1}, u_t)) \quad (9)$$

$$p_\theta(x_t|z_t, s_t) = N(x; \mu_x(z_t, s_t), \sigma_x(z_t, s_t)) \quad (10)$$

$$p_\theta(y_t|z_t, r_t) = N(y; \mu_y(z_t, r_t), \sigma_y(z_t, r_t)) \quad (11)$$

式 (9)、式 (10)、式 (11) はそれぞれマルコフ性を仮定し

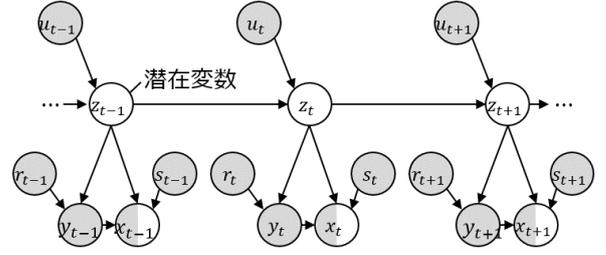


図 2 SDSSM のグラフィカルモデル

た状態モデル、観測モデル、観測モデルであり、[6]において z_t は潜在変数、 r_t と s_t 、 u_t 、 y_t は温度や湿度など連続データを得ることが可能な観測変数、 x_t は糖度など連続データを得ることが可能でない、または、連続データを得るためのコストが高い観測変数を仮定している。また、観測変数 x_t と y_t で潜在変数 z_t を共有することで、 x_t が欠損していた場合であっても、 y_t を用いて潜在変数の更新を可能とする。ここで、 μ_z と σ_z 、 μ_x 、 σ_x 、 μ_y 、 σ_y は式 (12)、式 (13)、式 (14) で表され、Deep neural networks (DNN) を用いて最適化することで任意の非線形関数で表現を可能とする。

$$\mu_z = NN_z(z_{t-1}, u_t) \quad \log \sigma_z = NN_z(z_{t-1}, u_t), \quad (12)$$

$$\mu_x = NN_x(z_t, s_t) \quad \log \sigma_x = NN_x(z_t, s_t), \quad (13)$$

$$\mu_y = NN_y(z_t, r_t) \quad \log \sigma_y = NN_y(z_t, r_t), \quad (14)$$

式 (12)、式 (13)、式 (14) において、 x_t は欠損のある観測変数を仮定しており、入力されるデータセットは観測変数 x_t を含んでいるラベル付きデータセット D_l と x_t を含まないラベルなしデータセット D_u の 2 種類が存在する。そのため、ラベル付きデータセット D_l とラベルなしデータセット D_u それぞれに対し、近似事後分布を用い、パラメータ θ と ϕ を最適化する ELBO である、 $L^l(x, y; \theta, \phi)$ と $L^u(x, y; \theta, \phi)$ を導出する必要があり、以下の式 (15)、式 (16) で表される。

$$\begin{aligned} & \log p_\theta(x_{1:T}, y_{1:T}) \\ & \geq \int q_\phi(z_{1:T}|x_{1:T}, y_{1:T}) \frac{p_\theta(x_{1:T}, y_{1:T}, z_{1:T})}{q_\phi(z_{1:T}|x_{1:T}, y_{1:T})} dz_{1:T} \\ & = \sum_{t=1}^T E_{q_\phi(z_t)} [\log p_\theta(y_t|z_t) + \log p_\theta(x_t|y_t, z_t)] \\ & \quad - \beta D_{KL}(q_\phi(z_1) || p_\theta(z_1)) \end{aligned} \quad (15)$$

$$- \sum_{t=2}^T E_{q_\phi(z_{t-1})} [\beta D_{KL}(q_\phi(z_t) || p_\theta(z_t|z_{t-1}))]$$

$$= -L^l(x, y; \theta, \phi)$$

$$\begin{aligned} & \log p_\theta(y_{1:T}) \\ & \geq \iint q_\phi(z_{1:T}, x_{1:T}|y_{1:T}) \frac{p_\theta(x_{1:T}, y_{1:T}, z_{1:T})}{q_\phi(z_{1:T}, x_{1:T}|y_{1:T})} dz_{1:T} dx_{1:T} \\ & = \sum_{t=1}^T E_{q_\phi(x_t|y_t)} [-L^l(x, y; \theta, \phi) + H[q_\phi(x_t|y_t)]] \end{aligned} \quad (16)$$

$$= -L^u(x, y; \theta, \phi)$$

式 (16) 内 $H[q_\varphi(x_t|y_t)]$ は近似事後分布 $q_\varphi(x_t|y_t)$ のエントロピーであり, $q_\varphi(x_t|y_t)$ の期待値である. ここで, ラベル付きデータセットの ELBO とラベルなしデータセットの ELBO を用いることで, モデルの学習に用いられる目的関数は式 (17) で表され, パラメータ θ と φ の最適化が行われる.

$$J = \sum_{D_l} L^l(x, y; \theta, \varphi) + \sum_{D_u} L^u(x, y; \theta, \varphi) + \alpha E_{D_u}[-\log q_\varphi(x_t|y_t)] \quad (17)$$

4. 基礎評価

4.1 概要

本節では半教師あり学習手法である SDSSM を用いた植物生育モデリング手法の有効性を示すために実施した評価について述べる. 半教師あり学習手法である SDSSM を用いた植物生育モデリング手法の有効性を示すため, SDSSM と代表的な深層学習手法である多層パーセプトロン (MLP), 時系列データの長期的な依存性を考慮可能な手法である Stacked Long short-term memory (sLSTM) [15] を用いて植物の積算蒸発散量推定精度を比較した. また, SDSSM はラベルデータ x_t が欠損している場合, ラベルなしデータセットを用いて潜在変数 z_t の更新を行う半教師あり学習モデル化手法である. そこで, SDSSM の半教師あり学習性能を検証するため, 学習時のラベルデータを間引いたデータセットを作成し, 間引き率に応じた推定性能比較を実施した.

植物糖度生育状態の指標である果実糖度は膨大な量のデータ収集が困難であるため, 代わりに, 積算蒸発散量を推定対象として用いることで, ラベルデータを間引いた際の推定精度が検証可能であり, SDSSM の半教師あり学習性能が検証できると考え, 本実験では推定対象に積算蒸発散量を用いた. また, MLP は入力層と出力層の他に隠れ層と呼ばれる中間層と呼ばれる層を持つことで, 任意の連続関数を近似可能な手法である. sLSTM は LSTM を多層化したものであり, 従来の LSTM に比べより長期的な依存性を考慮可能な手法である. そのため, 従来手法に比べ, 時系列データの長期的な依存性を考慮しつつ, 高精度な推定が可能であることを示すため, 基礎評価として MLP, sLSTM を用いた比較評価を実施した. 本実験における MLP と sLSTM, SDSSM の実装には Chainer[16]を用い, CPU Intel Core i7-6700K, GPU NVIDIA Geforce GTX1080 を搭載した PC で実施した.

精度比較では回帰問題における誤差指標として一般的である平均絶対誤差 (MAE), 二乗平均平方誤差 (RMSE) を用いた. 各指標の算出式を以下に示す.

$$MAE = \frac{1}{N} \sum_i^N |y_i - \hat{y}_i| \quad (18)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_i^N (y_i - \hat{y}_i)^2} \quad (19)$$

N はデータ数, y_i と \hat{y}_i はそれぞれ時点 i の真値と推定値, \bar{y} は真値の平均値である. MAE と RMSE はそれぞれ誤差指標であり, 値が小さいほど高精度に推定できたといえる.

4.2 データセット

本節では評価に使用したデータセットについて述べる. データセットには, 静岡大学農学部実験圃場における高糖度トマト低段密植栽培内で収集した実データを用いた. データセットの内容を表 1 に, 栽培データ収集環境図を図 3 に示す. 実験用植木鉢に植えられたトマト株に対し, 無線散乱光センサノード[17]と茎径計測用レーザ変位センサ (HL-T1, Panasonic 製) を設置し, 実験用植木鉢を電子天秤 (EK-300i, A&D 製) の上に設置した. 図 3 に示した栽培データ収集環境を図 4 に示す栽培実験圃場の 8 箇所に設置し, データ収集を実施した. 無線散乱光センサノードでは温度と相対湿度, 光量を経時計測した, 茎径センサデータでは茎径を経時計測, 電子天秤では株重量を経時計測した. また, 飽差は無線散乱光センサノードで計測した温度と相対湿度から算出した値であり, 一定の空気中に含むことのできる水蒸気量を表し, 蒸発散のしやすさに関係する. ここで, 糖度は膨大な量のデータ収集が困難であるため, 代わりに, 膨大な量のデータ収集が可能な積算蒸発散量を推定対象として用いることで, ラベルデータを間引いた際の推定精度が検証可能であり, SDSSM の半教師あり学習性能が検証できると考え, 本実験では電子天秤で測定した株重量から算出した値である, 積算蒸発散量を目的変数とした.

4.3 観測変数の検討

本節では SDSSM を用いた植物モデル化時の各観測変数の検討について述べる. SDSSM では潜在変数 z_t 以外の観測変数 (x_t, y_t, r_t, s_t, u_t) への説明変数を割り当てるかの設定を行う必要がある. 本実験における観測変数の内訳を表 2 に示す. 本実験では目的変数には積算蒸発散量を用いるため, 観測変数 x_t は積算蒸発散量である. ここで, 本稿における積算蒸発散量とは前回灌水時からの積算蒸発散量を示す. また, 一般に植物の蒸散は太陽光放射から受ける熱を分散させる手段のひとつであり, 気孔が開くことで行われると言われている. 同様に, 蒸散は葉の内部の水蒸気濃度と外部の空気の水蒸気濃度との差によって, 濃度の高い方から低い方へ水分子が拡散されることで行われる. そのため, 観測変数 x_t に作用する観測変数 s_t は光量と飽差, 灌水後

表 1 データセット内容

項目	内容
目的変数	積算蒸発散量[g]
説明変数	温度[°C], 相対湿度[%], 光量[V], 飽差[kPa], 茎径[V], 灌水後の経過時間[分]
データ周期	2 分
データ収集期間	2017/05/18 ~ 2017/07/24
データ収集時間	7:00 ~ 19:00

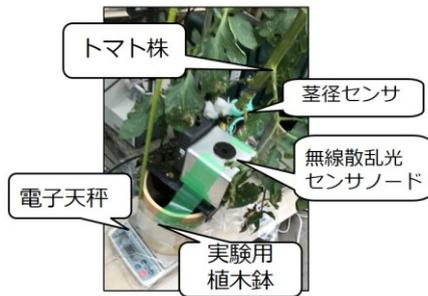


図 3 栽培データ収集環境図

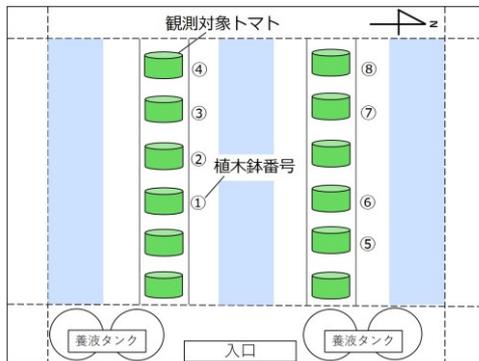


図 4 栽培実験圃場

の経過時間とした。植物の水分ストレスは茎径と密接な関係があることが知られており、植物の茎は水分ストレスの付与に伴い縮小する[18, 19]。したがって、茎径は間接的に植物状態を示しており、積算蒸発散量と潜在変数 z_t を共有可能であると考え、観測変数 y_t は茎径とした。また、茎径の変化は蒸散が行われ、植物の根から水分を茎や葉に押し上げる性質に強く依存すると考えられる。そのため、観測変数 r_t は積算蒸発散量に作用する観測変数 s_t と同様の光量と飽差、灌水後の経過時間とした。植物状態である潜在変数 z_t に作用する観測変数 u_t へは環境データである温度と相対湿度、光量を用いた。

4.4 SDSSM による積算蒸発散量推定性能比較

既存手法である MLP と sLSTM での積算蒸発散量推定精

表 2 観測変数内訳

観測変数	データ種類
x_t	積算蒸発散量[g]
y_t	茎径[V]
r_t	光量[V], 飽差[kPa], 灌水後の経過時間[分]
s_t	光量[V], 飽差[kPa], 灌水後の経過時間[分]
u_t	温度[°C], 相対湿度[%], 光量[V]

表 3 訓練・検証・評価データの内訳

データ種類	実験用植木鉢番号	データ数
訓練	1, 3, 4, 6, 7	48,600
検証	5, 8	19,440
評価	2	9,720

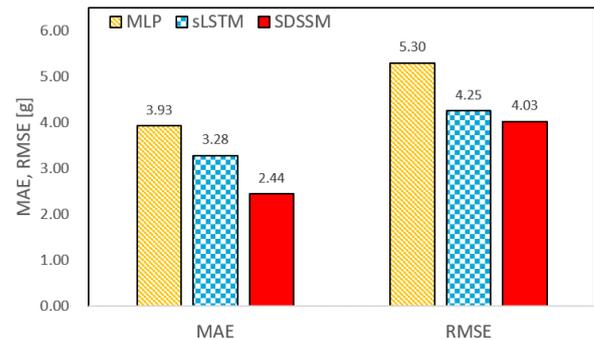


図 5 既存手法と提案手法の推定誤差

度と SDSSM を用いた積算蒸発散量推定精度を比較する。それぞれの手法はデータセットを訓練データと検証データ、評価データに分け、訓練、検証データでモデルの学習を行い、評価データを用いて推定精度を比較した。表 3 に訓練データと検証データ、評価データの内訳を示す。また、MLP と LSTM の説明変数には表 1 内の説明変数を用いた。既存手法と提案手法の推定誤差として、MAE と RMSE を図 5 に示す。提案手法の MAE は 2.44 と推定誤差が低く、MLP に比べ約 37.9%、sLSTM に比べ約 25.6%の誤差を削減できた。RMSE においても提案手法は 4.03 と低く、MLP に比べ約 24%、sLSTM に比べ約 5.3%の誤差を削減できた。ここで、図 7 に評価データにおける 2017 年 6 月 11 日の真値と SDSSM, sLSTM の推定値の推移を示す。図 7 から sLSTM は灌水からの時刻が 0 であるときに約 1、灌水からの時刻が 0 以外のときに約 10 を出力するという過学習をしていると考えられる。一方、SDSSM は過学習せず積算蒸発散量を追従できていることがわかる。そのため、SDSSM は過学習が起きにくい頑健なモデルであることが明らかとなり、

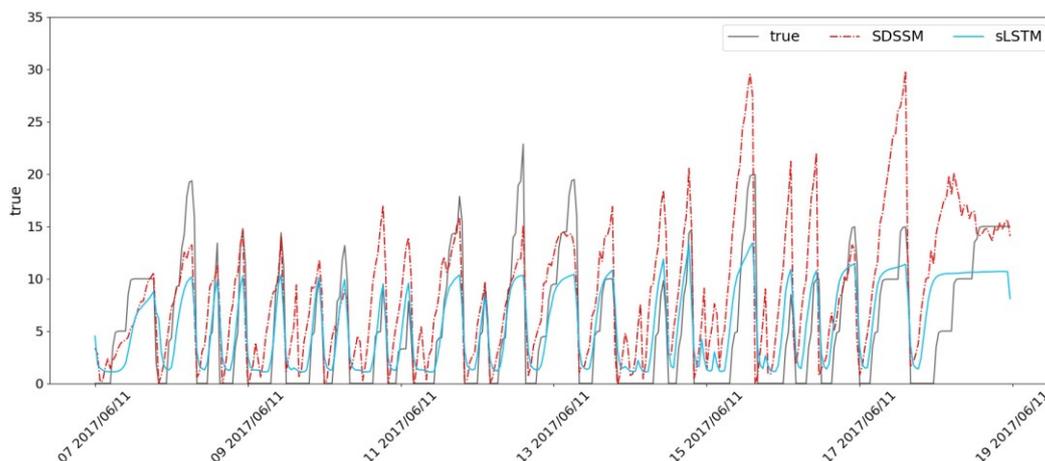


図 7 真値と推定値の推移

SDSSM を用いて作成した植物の生育モデルは RL における環境やシミュレータへの応用が期待できる。

4.5 学習データ数に応じた SDSSM 性能比較

SDSSM はラベルデータ x_t が欠損している場合、ラベルなしデータセットを用いて潜在変数 z_t の更新を行う半教師あり学習モデル化手法である。そこで、SDSSM の推定精度と学習時のラベルデータ数の関係を分析するため、学習時のラベルデータを間引くことでデータ数を削減した場合の推定精度を分析した。訓練データと検証データ、評価データの分け方は表 3 に示す。ラベルデータの間引き率は 0%, 25%, 50%, 75% とし、植木鉢ひとつ当たりのデータを連続した 4 ブロックに分割し、ブロック群を作成し、間引き率 25% ではブロック群における先頭の 1 ブロック、間引き率 50% では先頭と 3 番目のブロック、間引き率 75% では先頭の 3 ブロックを単純に間引いた。また、最後のブロック群における 4 番目のブロックは間引かずラベルデータとした。植木鉢ひとつ当たりの間引き率とラベルデータ数を表 4 に示し、ラベルデータの間引き率に対する推定誤差を図 6 に示す。図 6 からラベルデータを 75% 間引いた MAE は 3.16, RMSE は 4.87 であり、ラベルデータの間引き率 0% における MAE は 2.44, RMSE は 4.03 である。SDSSM では間引き率の増加、ラベルデータの減少に伴い、真値との誤差が大きくなる傾向があるが、低い推定誤差を維持していることがわかる。したがって、SDSSM ではラベルデータが欠損している箇所が存在しても、潜在変数を共有する別の観測データを用いて植物生育モデリングが行えることが明らかになった。SDSSM を植物生育モデリングに用いることで、糖度のような収集コストが甚大であり、大規模なデータ数の確保が困難なデータに対し、積算蒸発散量と茎径のようにデータから得られる潜在変数が類似しているデータしていると考えられるデータを用いることで、データの欠損を補い、高精度な推定が期待できる。

表 4 植木鉢ひとつ当たりの間引き率とラベルデータ数

間引き率 [%]	植木鉢ひとつ当たりのデータ数 [件]	訓練データ数 [件]	検証データ数 [件]
0	9720	48,600	19,440
25	7200	36,000	14,400
50	5040	25,200	10,080
75	2520	12,600	5040

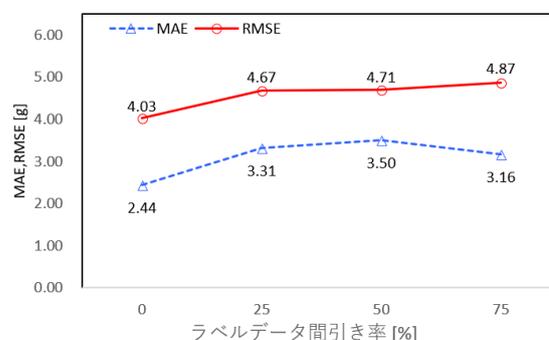


図 6 ラベルデータの間引き率に対する推定誤差

5. おわりに

本研究では、膨大な量のデータ収集が困難な糖度データを用いた植物生育モデリングを実現するため、半教師あり学習モデル化手法である SDSSM を用いた植物生育モデリング手法を検討した。トマト栽培時に収集した実データによる基礎評価を行った結果、提案手法は深層学習の代表的な手法である MLP に比べ MAE で約 37.9%, ネットワーク内部に再帰構造を持ち、長・短期記憶が可能な LSTM を多層化した sLSTM に比べ、MAE で約 25.6% の誤差を削減で

きることを確認した。また、訓練・検証データのラベルデータを間引き、間引き率に応じた SDSSM 推定精度の分析を行った結果、ラベルデータが約 75%欠損していても、SDSSM は潜在変数を共有する別の観測変数を用いて潜在変数を更新し、推定精度を維持することを確認した。そのため、農業分野をはじめとした大規模なデータ収集コストが大きい分野において有効な手法であると考えられる。今後、糖度データを用いた半教師あり学習性能の比較を行うため、実データの拡充を進めるとともに、植物モデリング時に画像データから得られる萎れ特徴量[20]を SDSSM への重畳手法の検討を行い、更なる植物生育モデリング精度の向上を目指す。

謝辞

本研究は JST さきがけ (JPMJPR15O5) の支援を受け実施されたものである。また、栽培データ収集環境をご提供いただいた静岡大学農学部の鈴木教授、鈴木様、澤村様に深い感謝の意を表す。

参考文献

- [1] 農林水産省：農村の現状に関する統計，農林水産省（オンライン），入手先<<http://www.maff.go.jp/j/tokei/sihyo/data/12.html>>（参照:2018-02-13）。
- [2] 農研機構：青年による農業のイメージ評価，農研機構（オンライン），< <http://www.naro.affrc.go.jp/project/results/laboratory/narc/1995/narc95-2-215.html>>（参照 2018-05-01）。
- [3] 大石直紀：トマトの養液栽培における水分ストレスに応じた給液制御システムの開発（1）茎径変化による水分ストレスの非破壊評価，生物環境調節，Vol.40, No.1, pp.81-89（2002）。
- [4] Kaneda, Y., Shibata, S., and Mineno, H: Multi-modal sliding window-based support vector regression for predicting plant water stress, KNOSYS, Vol.134, pp.135-148(2017).
- [5] 若森和昌，柴田瞬，峰野博史：深層学習を用いた植物の水分ストレス推定手法の検討，DICOMO2017, pp.199-206（2017）。
- [6] Shibata, S., Mizuno, R., and Mineno, H.: Semi-supervised deep state-space model for plant growth modeling, (submitted to Knowledge-Based Systems)
- [7] Chen, D., Dawei, H., Yuming, F., et al.: Analysis and optimization of the effect of light and nutrient solution on wheat growth and development using an inverse system model strategy, Computes and Electronics in Agriculture, Vol.109, pp.221-231 (2014).
- [8] Martin, L., Lars, K., and Amy, L.: A Review of Unsupervised Feature Learning and Deep Learning for Time-Series Modeling, Pattern Recognition Letters, Vol.42, pp.11-24(2014).
- [9] Sam, R., Zoubin, G.: A Unifying Review of Linear Gaussian Models, Neural Computation, Vol.11, No.2(1999).
- [10] 深谷肇一：状態空間モデルによる時系列解析とその生態学への応用，日本生態学会誌，Vol.66, pp.375-389(2016).
- [11] Andrew, J: Stochastic processes and filtering theory, Courier corporation(2007).
- [12] Kingma, D., Welling, M.: Auto-Encodig Variational Bayes, arXiv preprint, arXiv: 1312.6114(2013).
- [13] Justin, B., Christian, O.: Learning Stochastic Recurrent Networks, arXiv preprint, arXiv: 1411.7610(2015).
- [14] Krishnan, R., Shalit, U., and Riedmiller, M.: Deep Kalman filters, arXiv preprint arXiv: 1511.05121(2015).
- [15] Sepp, H., Jurgen, S.: Long-short-term memory, Neural Computation, Vol.9, No.8(1997).
- [16] Tokui, S., Oono, K., Hido, S., et al.: Chainer: a next-generation open source framework for deep learning, Proc. workshop on machine learning systems in the twenty-ninth annual conference on neural information processing systems, pp.1-6(2015).
- [17] Ibayashi, H., Kaneda, Y., Mineno, H., et al.: A Reliable Wireless Control System for Tomato Hydroponics, MDPI Sensors, Vol.16(2016).
- [18] Wang, X., Meng, Z., Lv, M., et al.: Determination of a suitable indicator of tomato water content based on stem diameter variation, Scientia Horticulturae, Vol.215, No.267, pp.142-148(2017).
- [19] Meng, Z., Duan, A. and Chen, D.: Suitable indicators using stem diameter variation-derived indices to monitor the water status of greenhouse tomato plants, Plos One 12.2(2017).
- [20] 柴田瞬，峰野博史：Optical Flowを用いた植物萎れ具合の推定，DICOMO2016, pp.66-72（2016）。