JSDSS 日本ソーシャルデータサイエンス論文誌
第 2 巻　第 1 号（2018 年 3 月）

特集：知的データサイエンス

機械学習と統計モデリングを組み合わせた手法による アスパラガスの収穫量推定

奥野 源[†1]　新谷 俊了[†1]

概要：昨今，ビジネスの様々な場面に機械学習を活用しようとする動きが多くみられている．その中でも，ディープラーニングと呼ばれる多層構造のニューラルネットワークを用いた機械学習が，画像認識・音声認識等の様々な分野で大きな成果を上げており，多くの実用例も紹介されている．しかしながら，ディープラーニング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という根拠を明示することが難しい，という弱点を持っている．このため，人の意思決定のサポートツールとして機械学習を利用する場合など，結果の根拠が重要となる場面では，別の手法を利用する必要がある．本記事では，このような場面に利用できる手法として，確率的因果推論の手法であるベイジアンネットワークと推定結果の直感的解釈が容易な重回帰モデリングを組み合わせた方法を，アスパラガスの収穫量推定に適用した事例を通じて紹介する．

キーワード：収穫量推定，機械学習，統計モデリング

1. はじめに

内閣府戦略的イノベーション創造プログラム (SIP)「農業データ連携基盤 (データプラットフォーム)」の構築事業に代表されるような昨今の農業分野でも，IoT の広まりにより圃場センサ等を用いてデータ収集・可視化する取り組みが広まっており，データ利活用の準備が整いつつある [1, 4].

この流れを受けて，本稿では土壌センサデータからアスパラガスの収穫量推定を試みる．農作物の将来の収穫量を事前に推定することは，農作業や出荷行動のマネジメントのためには非常に重要となる．例えば，翌週の作物の収穫量が大きく増加することが推定できれば，その週は近所の農家から人員の協力を仰ぐといった対策を，前もって準備することができる．

データを活用して推論モデルを構築する方法としては，ディープラーニングと呼ばれる多層構造のニューラルネットワークを用いた機械学習が，最近画像認識・音声認識などの様々な分野で大きな成果を上げており，注目が集まっている．しかしながら，ディープラーニング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指摘されている．また大前提として，ディープラーニングの有用性は，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し，データ量が十分でないと高い性能を発揮できないとされている [5]. これらの点は，多くの場面でビジネス応用における課題となり得る．

本稿では，アスパラガスの収穫量推定モデルを将来的に農家の方向けの意思決定のサポートツールとして活用していくことを目標に据え，やや小規模なデータからでもモデルが構築でき，かつどうしてその推定結果が得られたのかという根拠を提示できる推定モデルの構築方法を検討する．

2. 検証用データと収穫量推定モデル構築方法

2.1 検証用データ

本検証に利用したデータは，山形県最上郡最上町のアスパラガス圃場 1 か所に設置された土壌センサから取得された圃場データと，農家の方が記録したその圃場で収穫されたアスパラガスの収穫量データである．データの収集期間は 1 シーズン（2016 年 08 月 12 日～2016 年 10 月 02 日の期間の 32 日分）となっている．

2.1.1 圃場データ

圃場データは，気温，湿度，日照量，土壌水分量，土壌 EC 値（土壌の養分量を表す指標）の情報が毎日 10 分毎に計測されている．

2.1.2 収穫量データ

収穫量データは，作物の品質（規格内，規格外）と収穫時間帯（午前，午後）の別に 4 クラスに分類されて記録されている．

図 1　最上町におけるアスパラガス栽培の様子

†1 （株）システム計画研究所/ISP　（連絡先：okuno@isp.co.jp）

2.2 問題設定

推定したい対象として，農家の方にとっての実効性と汎用性の観点から，規格内アスパラガスの翌日収穫量増減率を据え，これを当日までの圃場データを使って推定することを目指す．

- 実効性
  農家の方が第一に気に掛けるのは，出荷対象となる規格内作物の収穫量であるため，これを対象にする．
- 汎用性
  作成した収穫量推定モデルを他の圃場にも展開していくことを考え，圃場の広さに相関があると思われる収穫量の値自体を推定対象にするのではなく，増減率を対象とする．

2.3 推定モデル構築方法

アスパラガスの収穫量は気象データによる重回帰モデルで簡易的に予測できることが分かっているので [3]，本記事でも推定モデルとしては重回帰モデルを採用した．その際，重回帰モデルをより現実の因果関係を近似したものにするため，予めベイジアンネットワークを使って変数間の因果関係を把握し，その因果構造を使ってモデルを生成した．

2.3.1 入力統計量の作成

推定モデルの入力データは，土壌センサから計測される圃場データとなるが，このデータは時間解像度が高く，そのままでは今回設定した推定対象（収穫量増減率）と大きく時間スケールが異なっている．このため，原系列データから積算量や上昇・下降トレンドを表す統計量を算出し，それを推定モデルの入力値として利用することとした．

2.3.2 因果関係のある要素の抽出

翌日収穫量増減率を正しく推定するためには，圃場データから作成された統計量の中から，アスパラガスの収穫量の増減に因果関係のある統計量を抽出し，その要素を使って推定モデルを作成する必要がある．そこで，確率的な因果関係を抽出することのできる手法としてベイジアンネットワークを採用し，この手法によって，アスパラガスの収穫量の増減に直接的な確率的因果関係のある統計量を抽出した．

2.3.3 変数選択とモデリング

抽出した統計量を使って収穫量増減率を推定するため，重回帰モデルを利用して収穫量推定モデルを構築した．その際，ベイジアンネットワークで抽出したすべての統計量を説明変数として回帰モデルに組み込んでしまうと，説明変数の数が多すぎて過学習を起こしてしまうことが分かった．そのため説明変数を絞り込む必要があるので，一般的な統計的変数選択手法である赤池情報量基準 (AIC) に基づくステップワイズ法と L1 および L2 正則化を同時に考慮した Elastic Net による方法を試行し，この結果を比較して最終的な収穫量推定モデルを構築することとした．

3. アスパラガスの収穫量推定

3.1 入力統計量作成

圃場データは，毎日 10 分毎に計測されているため，一日当たり 144 個の時系列データが得られている．この時間解像度の高い原系列を何らかの方法で集約した統計量を作成するが，これは解釈可能なものである必要がある．なぜならば，この統計量が推定モデルの入力であり，すなわち推定結果の根拠を表す量になる．したがって，意思決定のサポートという目的のためには，この統計量を解釈のつけられるものとすることが重要となる．

本記事では，気温・湿度・日照量の気象に関する 3 項目については，日ごとの積算量を利用することとした．これは，植物に対して広く論じられている有効積算温度のアイデア [2, 6] から，気象に関する情報を積算することで，植物の生育に対する因果を導けるのではないかと考えたからである．一方，土壌水分量と土壌 EC 値の土壌に関する 2 項目については，謝辞でも言及している最上町の農家の方から得られた経験則である「雨後は急激にアスパラガスが成長する」というノウハウの計量を目指し，日ごとの上昇・下降トレンドを算出した．

表 1　圃場データの集計方法

項目	集計方法	集計期間
気温	積算量	基準日（予測日前日）午前までの積算と基準日前日までの積算を利用した．
湿度		
日照量		
土壌水分量	上昇・下降トレンド	
土壌 EC 値		

ここで，上昇・下降トレンドは， 6 時間分の時間幅による原系列移動平均データの Kendall 順位相関係数（時系列昇順の順位との相関）が±0.7 を超えた回数として計量した．相関を取る対象を原系列ではなく移動平均データとした理由は，原系列の変動がかなり細かい時間的粒度で発生していることにある．この変動は，土壌センサの計測誤差等に起因するものと考えられ，大域的変動傾向をつかむためには，予めこの誤差変動を丸める必要があると考え，移動平均後に順位相関を求めた．

また集計期間については，基準日（予測日前日）午前までの積算に関して，基準日のみ，前日から，2 日前から，3 日前からの 4 パターン，基準日前日までの積算に関して，前日のみ，2 日前から，3 日前からの 3 パターンとした．

3.2 因果関係のある要素の抽出

収穫量に対して確率的な因果が強く認められる統計量が収穫量推定モデルの説明変数として適切であると仮定し，ベイジアンネットワークを使って抽出した．ネットワーク作成時の時系列的制約条件として，時系列的に過去に向かう因果関係（例えば，過去 1 日分の積算量から過去 2 日分の積算量に向かう因果関係のようなもの）は禁止した．

ベイジアンネットワークの有向グラフにおいて，目的変数となる収穫量に対して他の統計量を介さずに直接連結された変数を抽出した結果は表 2 の通りである．なお．互いに強い相関（絶対値 0.8 以上）が認められた項目については，因果関係が弱い方を除去している．

表 2　ベイジアンネットワークで抽出された項目

項目	集計方法	集計期間
日照量	積算量	前日から午前まで
気温	積算量	前日のみ
気温	積算量	午前中のみ
土壌 EC 値	下降トレンド	3 日前から午前まで
土壌 EC 値	下降トレンド	2 日前から午前まで
土壌 EC 値	下降トレンド	午前中のみ
土壌 EC 値	上昇トレンド	前日のみ
土壌 EC 値	上昇トレンド	午前中のみ
土壌水分量	上昇トレンド	午前中のみ
土壌水分量	下降トレンド	前日から午前まで
土壌水分量	上昇トレンド	2 日前から午前まで
土壌 EC 値	下降トレンド	前日のみ
土壌 EC 値	上昇トレンド	前日から午前まで
土壌水分量	上昇トレンド	前日から午前まで
土壌水分量	下降トレンド	午前中のみ

3.3 変数選択とモデリング

抽出された統計量に対して，AIC によるステップワイズ法と Elastic Net による変数選択を実施し，重回帰モデルを構築し推定精度を比較したところ，AIC によるステップワイズ法による変数選択の方が高精度となった．

作成した 2 つの回帰モデルに対して 5-fold クロスバリデーション（5-fold CV）により精度を評価した結果を表 3，表 4 に示す．比較すると，5 回中 4 回のクロスバリデーションケースで AIC に基づくステップワイズ法による重回帰モデルの検証時推定誤差が小さくなった．回帰モデルに含まれる説明変数の数を比較すると Elastic Net によるものの方が多く，加えて学習時と検証時の推定誤差の差分も大きくなる傾向があることから，Elastic Net によるモデルは複雑すぎて，やや過学習気味になっていると考えられる．なお，5-fold CV は全 32 日分のデータの 4/5 を学習用データに，1/5 を検証用データに利用して行った．誤差の指標としては平均 2 乗誤差 (MSE) を用いた．

表 3　AIC によるステップワイズ法で構成した重回帰モデルに対する 5-fold CV の結果

CV#	学習		検証	
	データ数	誤差	データ数	誤差
1	25	0.026	7	0.035
2	25	0.032	7	0.015
3	26	0.026	6	0.031
4	26	0.026	6	0.039
5	26	0.020	6	0.071

表 4　Elastic Net による変数選択によって構成した重回帰モデルに対する 5-fold CV の結果

CV#	学習		検証	
	データ数	誤差	データ数	誤差
1	25	0.021	7	0.081
2	25	0.030	7	0.022
3	26	0.024	6	0.045
4	26	0.022	6	0.052
5	26	0.020	6	0.067

5-fold CV 検証時に推定誤差が小さくなった AIC によるステップワイズ法によって選択された項目は，表 5 の通りである．

表 5　AIC によるステップワイズ法で選択された項目

項目	集計方法	集計期間
日照量	積算量	前日から午前まで
気温	積算量	前日のみ
土壌 EC 値	上昇トレンド	前日のみ
土壌水分量	上昇トレンド	2 日前から午前まで
土壌 EC 値	下降トレンド	前日のみ

その際の 5-fold CV で作成された推定モデルの各係数値は，表 6 の通りである．

5 回の結果を比較すると，ほぼ同じような係数の傾向を持ったモデルが作成されていることから，学習データに対する大きな偏りはなかったと考えられる．そこで，この 5 つの推定モデルの各係数の平均を求め，その平均値を利用した収穫量推定モデル（平均モデル）を作成し，最終的な推定精度を評価した．平均モデルにより全 32 日分の圃場データから収穫量増減率を推定した結果を図 2 に示す．

表 6　5-fold CV の各モデルの係数

項目	CV#1	CV#2	CV#3	CV#4	CV#5
切片	0.978	0.980	0.957	0.945	0.978
日照量 積算量	0.228	0.270	0.259	0.248	0.244
気温 積算量	-0.116	-0.136	-0.110	-0.092	-0.112
土壌 EC 値 上昇トレンド	-0.041	-0.072	-0.074	-0.052	-0.126
土壌水分量 上昇トレンド	-0.086	-0.096	-0.077	-0.094	-0.029
土壌 EC 値 下降トレンド	0.131	0.138	0.115	0.124	0.100

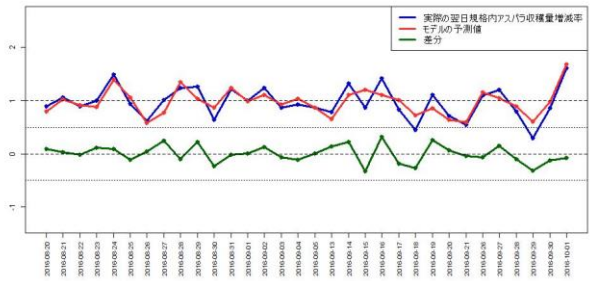

図 2　収穫量増減率推定の結果

推定結果の MSE は 0.027 となった．実際の収穫量増減率（青線）と推定結果（赤線）を比較すると，増加・減少の傾向をとらえた推定ができていることが分かる．

4. 推定モデルに対する解釈

重回帰モデルの場合，説明変数の係数は陽に求まるため，解釈を行うことができる．なお，説明変数とした各項目値は標準化を行っているため，係数値の絶対値が大きいほど収穫量の増加に強い影響を与えていると考えることができる．また，係数値の正負には意味があり，係数値正の場合は項目値が大きいほど，負の場合は小さいほど，収穫量の増加に強く影響すると解釈できる．

3.1 節で作成した各統計量の相関を求めたところ，1 日分の気温積算量と湿度積算量の間には負の相関（相関係数-0.74）が観測されているので，気温が低いと湿度が高い，すなわち天候は雨の可能性が高かろうと推定することができる．したがって，推定モデルから読み取れる前日の気温が低いと収穫量が増えるという解釈は，雨後はアスパラガスが急成長するという最上町のアスパラ農家さんのノウハウを間接的に表していると考えることができる．

表 7　推定モデルに対する解釈

項目	集計期間	係数値	収穫量増加要因
日照量 積算量	前日から 午前まで	0.2500	前日から日照量が多い
気温 積算量	前日のみ	-0.1133	前日の気温が低い
土壌 EC 値 上昇トレンド	前日のみ	-0.0732	前日の土壌 EC 値上昇が少ない
土壌水分量 上昇トレンド	2 日前から 午前まで	-0.0762	2 日前からの土壌水分量上昇が少ない
土壌 EC 値 下降トレンド	前日のみ	0.1214	前日の土壌 EC 値下降が多い

それ以外にも，日照量が多い場合に収穫量が増えるというのは，光合成の影響が考えられるので直感的に納得感が高い．また，土壌 EC 値（土壌の養分量を表す指標）の下降傾向は，養分が植物の生育により消費された結果だと考えることで，これが結果的に収穫量増加につながったと解釈できる．

5. まとめと課題

5.1 まとめ

本稿で提案した機械学習と統計モデリングを組み合わせた手法によって，推定結果に明確な根拠を示すことのできる回帰モデルを生成し，ある程度傾向を捉えた収穫量推定を行うことができるということが検証できた．生成されたモデルには，直感的に納得感の高いパラメータ（正係数の積算日照量など）や農家の方のノウハウを間接的に表現したパラメータ（負係数の積算気温）が組み込まれており，実際の複雑な植物生育プロセスの因果関係を部分的にでも表現できたモデルになっているのではないかと考えている．

4 節では土壌 EC 値（土壌の養分量を表す指標）の下降傾向を土壌養分が植物の生育により消費された結果だと考えることで収穫量増加につながったと解釈した．もしこの仮説が正しいとすると，土壌 EC 値については追肥によってある程度コントロールできると想定されるので，本検証のビジネス応用として，土壌 EC 値の傾向を用いた農家の方への追肥行動リコメンドといったことも見えてくるのではないだろうか．

5.2 課題

今回取り扱ったデータは，1 農家 1 シーズンのアスパラガス栽培データであった．モデルの精度検証時には，5-fold CV を実施することにより過学習を起こしていないことの検証を行ったが，そもそもクロスバリデーションで利用したデータが，非常に限定された条件におけるバリエーションの少ないデータであるため，より汎用的な場面でも実効

性のあるモデルが作成されたか否かは検証できていない．

加えて今回作成した収穫量予測の回帰モデルには，筆者には直感的な説明がつきにくいパラメータも含まれていた．これらについては，農家の方に実運用の中で評価してもらい，モデリングプロセスに起因する不適切なパラメータであるか否かを見極める必要がある．

したがって，今後実運用を考える際には，別のシーズン・圃場のデータを収集し追加検証を行うとともに，実際に農家の方に利用・評価いただき，フィードバックを重ねていく必要がある．

また，今回は既存研究 [3] を参考に重回帰モデルによる推定を行ったが，ベイジアンネットワークを利用すると因果構造の把握だけでなく，推論まで行うことができるため，どちらがより適切な推定手法となるかという検証も，今後の課題として考えている．

6. 今後の機械学習への期待

いま世間では，機械学習に対する過度な期待が広まっているように感じられる．曰く，機械学習，特にディープラーニングは万能でどんな課題でも解決でき，やがて人間の仕事のほとんどは，これを活用したシステムに置き換わる，といったようなものである．

確かに機械学習は，正しく活用すれば非常に強力なデータ利活用ツールの一つとなり得，ルールベースのアプローチでは気づくことのできなかった知見を得ることができる可能性を秘めている．しかしながら機械学習の本質は，学習したパターンに基づいて類似パターンを探し出すといったものであり，まったく未知の問題に対する予測ができるものではない．つまり，複雑な社会における人間の意思決定プロセスを，すべて肩代わりできるものではないはずである．

筆者は機械学習を，困りごとを丸投げできるようなある種の人間の代替としてではなく，意思決定の材料を提供してくれる賢いパートナーのように活用することで，非常に高いパフォーマンスを発揮するツールだと思っている．機械学習を期待外れな一過性のブームにせず，本当に社会的価値のあるツールとして浸透させていくためには，活用範囲や用途を見誤らず，適切な形で社会実装していくことが求められている．

謝辞

アスパラガス栽培に関するノウハウについては，2016 年 3 月 3 日に山形県最上郡で開催された「最上町 農業 x IT アイデアワーク」に於いて，最上町でアスパラガスを栽培されている農家の方をはじめ，たくさんの方から助言をいただいた．また，日本ソーシャルデータサイエンス学会事務局の皆さまには，本記事を発表する機会をいただいた．ここに記して感謝いたします．

図 3　最上町 農業 x IT アイデアワークの様子

参考文献

[1] 大臣官房政策課技術政策室,「農業データ連携基盤の構築について」, (2017). http://www.affrc.maff.go.jp/docs/genba/attach/pdf/sanko1.pdf （2018 年 01 月 18 日アクセス）

[2] 林三徳，伏原肇，柴戸靖志，“有効積算温度法による軟弱野菜の収穫期の予測,” 平成 3～5 年度園芸研究所野菜花き部野菜品種研究室試験成績書 (1993).

[3] 岸田史生，榊原俊雄，黒澤俊，西川浩次，楠見浩二，札埜高志，片岡圭子，北島宣，“公開気象観測データを利用したグリーンアスパラガスの日別収量の簡易予測,” 農業生産技術管理学会誌 Vol. 20, No.3, pp. 79-84 (2013).

[4] 金間大介，野村稔，“農業をめぐる IT 化の動き―データ収集、処理、クラウドサービスの適用事例を中心に―,” 科学技術動向，No. 142, pp. 13-18 (2014).

[5] Marcus, G. *Deep Learning: A Critical Appraisal.* arXiv: 1801.00631 Cornell University Library (2018).

[6] 岡正明，大山優美子，小川貴史，“有効積算温度を用いたエダマメ品種の収穫適期予測法,” 宮城教育大学紀要，Vol. 40, pp.201-208 (2005).