F-043

時系列パターンの共起性に基づく大豆の収量に関与する土壌水分環境の抽出
Extraction of Soil Moisture Environment Affecting Soybean Yield Based on Co-occurrence of Time Series Patterns

逸見 聡* 東山 久瑠実* 長南 友也† 林 怜史† 中村 卓司†
Satoshi Henmi Kurumi Higashiyama Yuya Chonan Satoshi Hayashi Takuji Nakamura
辻 博之† 村上 則幸† 西出 亮‡ 大川 剛直* 小澤 誠一§
Hiroyuki Tsuji Noriyuki Murakami Ryo Nishide Takenao Ohkawa Seiichi Ozawa

1. はじめに

近年，日本国内では農業従事者の高齢化や後継者不足が問題となっている．そのため，科学的方法を用いた技術継承の支援や収量の向上，農作業の効率化が求められている．このような課題に対して，蓄積された膨大なデータを基にデータマイニングによって，栽培に有用な知識・知見を発見する試みが見られる．

大豆の収量は近年伸び悩んでおり，諸外国と比べても低い水準となっている．これには多数の多収阻害要因が影響しているものと考えられる．そこで，農林水産省委託のもと，大豆や麦類などの作物を対象に多収阻害要因の解明を目的としたプロジェクト(2015 年度～2019 年度)が推進された．我々の研究グループもこのプロジェクトに参加し，大豆を取り巻く栽培環境に着目することで，多収あるいは低収となる要因の分析・把握を目的とした研究を行っている．

大豆を取り巻く栽培環境には，気象，土壌特性，土壌水分などがあり，中でも土壌水分や土壌特性の 1 つである排水性は大豆の生育に大きな影響を与えると考えられている．土壌の排水性は，降雨が発生したのちの土壌水分の時系列変化から捉えることができる． 例えば，降雨によって土壌水分が増加した後，1 日で土壌水分が降雨前に戻る土壌と，何日も土壌水分が増加したままである土壌とでは前者は排水性が良く，後者は排水性が悪いと判断できる．このように，類似した降雨が発生したのちの土壌水分の変化を見ることで，圃場ごとの土壌の排水性を捉えることができる．

我々は，様々な環境下に共通して現れる土壌水分時系列変化を抽出することで，収量に関与する土壌水分環境を発見する手法を提案している[1]．しかしながら，この手法では降水量を考慮していないため，土壌水分の変化が排水性など土壌の特性によるものであるか，雨水の供給によるものであるか明確ではなく，多収阻害要因分析の際に曖昧さが残っていた．

そこで本研究では，降水量と土壌水分の特徴的，あるいは一般的な変化をパターン化することで把握し，それらの共起に着目した分析を行うことで，多収あるいは低収に関与する排水性や土壌水分環境を発見する手法について検討する．なお，共起という言葉は自然言語処理の分野で用いられる用語であるが，本論文では 2 つの変化が同時に出現するという意味で用いる．

* 神戸大学大学院 システム情報学研究科
Kobe University Graduate School of System Informatics
† 北海道農業研究センター
NARO Hokkaido Agricultural Research Center
‡ 滋賀大学 データサイエンス教育研究センター
Shiga University,
The Center for Data Science Education and Research
§ 神戸大学 数理・データサイエンスセンター
Kobe University Center for Mathematical and Data Sciences

2. 実態調査データ

大豆の収量や品質の向上のために，土壌の成分や土壌水分，病虫害や雑草害などの把握が求められている．そこで，農林水産省委託の多収阻害要因の診断法及び対策技術の開発プロジェクト (2015 年度～2019 年度) では，センサによる測定や農家へのアンケートを行うことで大豆栽培の実態に関するデータを取得し，蓄積している．この取得されたデータには，大まかに分けて土壌水分データ，栽培環境データ，収量データがある．

- 土壌水分データ：圃場の土壌全体積のうち水分が占める割合を示すデータ．圃場ごとに一定間隔で繰り返し取得されるため，時間とともに変化する時系列データである．
- 栽培環境データ：圃場の緯度経度や作業・耕種に関して調査したデータ．これには，播種日や開花日，土壌の状態を示す 1 つの指標である仮比重の情報が含まれている．
- 収量データ：収量の指標の 1 つである，精子実重データ．これを参照することで圃場が多収あるいは低収であるかを判断できる．

以上のデータの他に，農研機構メッシュ農業データシステム[2]を通じて，緯度経度情報から圃場ごとの時系列降水量データを取得することができる．

3. 降水量と土壌水分を基にした収量に関与する土壌水分環境の発見

3.1 概要

本研究では，対象となる圃場ごとに取得された実態調査データを基に，多収あるいは低収に関与する土壌の排水性や土壌水分環境を把握する手法の開発を目的としている．データの取得されている圃場をあらかじめ多収圃場と低収圃場に分類し，それぞれの圃場に頻出する土壌水分の排水性や土壌水分環境の特徴を抽出することで多収要因，低収要因を発見する．提案手法の概要を図 1 に示す．

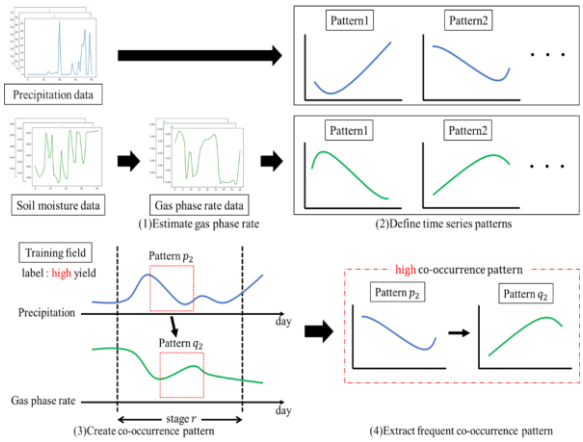

図 1 ：提案手法の概要

圃場ごとの土壌の排水性や土壌水分環境を定義する．排水性は，降水量と土壌水分の特徴的，あるいは一般的な時系列変化の相関を見ることで把握できる．時系列変化を定義するため，それぞれの時系列データを一定間隔で区切り，パターン化する．このようにしてパターン化した部分時系列を時系列パターンと呼ぶ．なお，土壌水分の変化は土壌特性によって挙動が大きく異なるため，異なる圃場間でも挙動の差が小さい土壌中の空気の割合である気相率に変換してパターン化することとする (図 1.(1)，図 1.(2)) ．

排水性を把握するため，降水量と気相率の共起を定義する．ここで，降水量の時系列パターンと気相率の時系列パターンの共起の組み合わせを共起パターンと呼ぶ．例えば，ある共起パターンは「一時的な降雨」という降水量の時系列パターンと，「減少後，急速に元の値に戻る」という気相率の時系列パターンの組み合わせを表す．このとき，どの期間における共起パターンが収量に関与するかがわからないため，組み合わせを網羅的に作成することが必要である．また，降水量の変化の後に，気相率の変化が起こり，両者の変化の間にはそこまでの時間を要さないことを考慮して，共起パターンは降水量の時系列パターンから一定期間内に見られる気相率の時系列パターンの組み合わせに限定する．得られた共起パターンは圃場ごとのある期間における土壌の排水性や土壌水分環境を表す (図 1.(3)) ．

収量に関与する共起パターンを抽出するため，圃場を多収圃場と低収圃場に分け，それぞれの圃場集合内で頻出する共起パターンを発見することを考える．これは，多収圃場集合，低収圃場集合をバスケットデータ，各圃場をトランザクション，共起パターンをアイテムとしたときの，系列パターンマイニング問題に帰着できる．系列パターンマイニングによって得られた多収圃場の頻出アイテム集合，低収圃場の頻出アイテム集合が，それぞれ多収要因の共起パターン，低収要因の共起パターンとなる．また，どちらの圃場にも頻出する共起パターンを考慮し，多収圃場集合，低収圃場集合の両方から発見された共起パターンは収量に関与しないとみなす (図 1.(4)) ．

発見された頻出アイテム集合を基に，収量が未知の圃場に対して収量を分類することを考える．対象の圃場がどれだけ多収要因の共起パターン，低収要因の共起パターンを，それぞれどれだけ持っているかを確認し，各共起パターンの割合に応じて多収低収のラベル付けを行う．

3.2 気相率の推定

土壌中の水分の挙動は，土壌の特性や状態によって大きく異なる．一方で，土壌中の空気の挙動は圃場間での差が小さい．このため，提案手法では土壌水分の値を用いるのではなく，土壌中の空気の割合を用いる．

土壌は気相 (空気)，液相 (水分)，固相 (土粒子)で構成され，これらの容積割合を気相率，液相率，固相率と呼ぶ．気相率をg，液相率をl，固相率sとすると式 (3.1) が成り立つ．

$$g + l + s = 1 \tag{3.1}$$

式 (3.1) を用いて，気相率を推定することを考える．液相率は土壌中の水分の容積割合であるため土壌水分と等価であり，固相率は仮比重と真比重で計算可能である．仮比重とは土壌のような多孔質物質の密度を表す値であり，真比重とは固相部分のみの比重を表す．仮比重をd_{bulk}，真比重をd_{pure}とすると，固相率sは式 (3.2) によって求められる．

$$s = \frac{d_{\mathrm{bulk}}}{d_{\mathrm{pure}}} \tag{3.2}$$

したがって，式 (3.1) と式 (3.2) より土壌水分を気相率gに変換することができる．気相率は，土壌水分が時系列データであるため時系列データである．なお，この気相率は推定した値であり，実際の値とは異なる．

3.3 時系列パターンの発見

大豆は生育段階ごとに望ましいとされる栽培環境が異なる．そこで，事前に時系列データを生育段階ごとに区切り，ぞれぞれの段階での栽培環境に着目することを考える．しかし，大豆の生育段階の捕捉には大豆の様子を継続的かつ詳細に観察する必要がある．そこで，開花日など生育段階を示す明確な基準日を用いながら，生育過程全体を一定期間に分割し疑似的な生育段階を仮定する．この疑似的な生育段階を生育ステージと呼ぶ．各時系列データに対して，開花日を基準とした生育ステージを定義し，ステージごとの栽培環境を抽出する．

土壌の排水性を把握するため，どのような降雨の後に，どのように気相率が変化したかを捉える．このために降水量，気相率の特徴的，あるいは一般的な時系列変化を捉えたい．しかし，そのような時系列変化がどのような時系列グラフで表されるかが事前に分からない．例えば，「降雨」という一般的な降水量の時系列変化でも，「大雨」や「小雨」，「断続的な降雨」や「ゲリラ豪雨」など様々な種類があり，具体的にどのような時系列グラフをしているのかがわからない．そこで，各時系列の部分時系列に対して，網羅的に特徴量ベースのクラスタリングを行い，得られたクラスタを時系列パターンとして定義する．なお，降水量と気相率では，時系列変化の仕方が異なるため，それぞれで時系列パターンを求める．

各圃場の時系列データを$x = \{x_1, x_2, \ldots, x_n\}$($n$:時系列長)とする．$x$に対して，一定期間で分割した生育ステージを定義し，$x = \{gs_1, gs_2, \ldots, gs_R\}, gs_r = \{x_1^r, x_2^r, \ldots, x_{\Delta t}^r\}$($gs$:生育ステージで区切られた時系列データ，$\Delta t$:各生育ステージの期間，$R$:生育ステージの数，$1 \le r \le R$，$\Delta t \times R = n$)とする．

次に，時系列変化を定義し，パターン化するため，xの部分時系列を網羅的にクラスタリングする．部分時系列を$sd_i^r = \{x_i^r, x_{i+1}^r, \ldots, x_{i+(L-1)}^r\}$($i$:サンプリングの開始インデックス，$L$:サンプリング長，$1 \le i \le \Delta t - (L-1)$)とし，各圃

場で定義したsd_i^rの集合に対して，特徴量ベースのクラスタリングを行う．ここで用いる特徴量は，sd_i^rを識別する特徴であり，最大値や最小値といった特徴を複数用いる．クラスタリングによって得られたクラスタ$C_k(1 \leq k \leq K$，K:クラスタ数)に属する部分時系列を，時系列データ内のk番目の時系列パターンとして定義する．

3.4 土壌の排水性を考慮した共起パターン

土壌の排水性を把握するため，どのタイミングで雨が降り，その後雨水がどのように排水されるかを把握する．このために降水量の時系列パターンと気相率の時系列パターンが，どのような生育段階のときに，どのような時間関係で共起しているかを把握することを考える．このとき，雨が降ってから気相率に変化が出るまで，それほど時間は掛からないと考えられる．そこで，適切な期間を定め，その期間内で発生する両時系列パターンの共起パターンを網羅的に定義する．

降水量と気相率の両時系列パターンを定義する．降水量の時系列データ$x = \{x_1, x_2, \dots, x_n\}$，気相率の時系列データ$y = \{y_1, y_2, \dots, y_n\}$に対して，3.3 節の手順で得られる降水量の時系列パターンデータを$p = \{gs_1, gs_2, \dots, gs_R\}, gs_r = \{p_1^r, p_2^r, \dots, p_{\Delta t-(L-1)}^r\}$，気相率の時系列パターンデータを$q = \{gs_1, gs_2, \dots, gs_R\}, gs_r = \{q_1^r, q_2^r, \dots, q_{\Delta t-(L-1)}^r\}$とする．なお，変数に用いられている添え字は 3.3 節と同様の意味を表す．また，各時系列パターンp_i^r，q_i^rには，対応する部分時系列sd_i^rが属するクラスタのインデックスが保持される．

降水量と気相率の両時系列パターンの組み合わせである共起パターンを定義する．このとき，一定期間内の様々な間隔の時間関係をとる両時系列パターンの組み合わせを網羅的に作成する．共起パターンデータ$OD = \{gs_1, gs_2, \dots, gs_R\}, gs_r = \{od_1^r, od_2^r, \dots, od_N^r\}, od_v^r = (p_{t_p}^r, q_{t_q}^r)$($N$:生育ステージ内でできる共起パターンの組み合わせの数，$1 \leq v \leq N$，t_p,t_q:各時系列パターンの開始インデックス)とする．t_p，t_qを式 (3.3) で定義する．

$$\begin{cases} t_p, t_q \in [1, \Delta t - (L-1)] \\ t_p \leq t_q \\ t_q - t_p \geq \alpha \\ \alpha \geq 0 \end{cases} \tag{3.3}$$

網羅的に共起パターンを定義するため，式 (3.3) を満たすすべてのt_p，t_qの組み合わせに対して$od_v^r = (p_{t_p}^r, q_{t_q}^r)$を定義する．また，$gs_r$の要素の順序は時系列順ではない．

以上の操作を行うことで，対象の圃場における生育ステージごとの排水性や土壌水分環境を表す共起パターンを保持した共起パターンデータが定義できる．

3.5 頻出アイテム集合に着目した，収量に関与する共起パターンの発見

多収圃場に頻出する特徴は多収要因，低収圃場に頻出する特徴は低収要因であると考えられる．そこで，圃場の降水量と気相率の共起パターンを保持しているデータODを基に，収量に関与する共起パターンを発見する．

多収圃場集合，低収圃場集合のそれぞれに対し，系列パターンマイニングアルゴリズム PrefixSpan[3]を適用することで，頻出アイテム集合，つまり各圃場集合に頻出する共起パターンを発見することができる．しかし，多収圃場と低収圃場の両方で同じ頻出アイテム集合が発見される可能性がある．これは，一般的に発生しやすい共起パターンが含まれているからである．そこで，多収の頻出アイテム集合を列挙したHF，低収の頻出アイテム集合を列挙したLFに対して，多収，低収の一方に稀有な頻出アイテム集合を列挙したHF'，LF'を式 (3.4) に従って得る．

$$\begin{gathered} F_1 \in HF, F_2 \in LF, \\ HF' = \{F_1 \mid F_1 \not\subseteq F_2\}, \\ LF' = \{F_2 \mid F_2 \not\subseteq F_1\} \end{gathered} \tag{3.4}$$

ここで，F_1，F_2は頻出アイテム集合を表す．以上より得られたHF'，LF'で列挙されている頻出アイテム集合が，多収，低収の一方に関与している共起パターンである．

3.6 収量が未知である圃場の分類

3.5 節より，多収，低収の一方に稀有な頻出アイテム集合が得られる．頻出アイテム集合を用いて収量が未知の圃場に対する分類を行うことを考える．なお，ここでは頻出アイテム集合を得るために用いた圃場の集合をトレーニングデータ，収量が未知の圃場の集合をテストデータと呼ぶ．

まず，テストデータの共起パターン$\widetilde{OD}$を作成するために，テストデータの降水量・気相率の時系列パターンを把握する．このとき，テストデータの時系列パターンはトレーニングデータとほぼ同様の時系列パターンで構成されると考えられるが，トレーニングデータと同様の方法で時系列パターンを定義しても，類似した時系列パターンが得られるとは限らない．そこで，トレーニングデータで構成された時系列パターンの中から，テストデータの部分時系列と最も類似している時系列パターンを探すことで，テストデータの降水量・気相率の時系列パターンを定義する．

部分時系列が時系列パターンと類似しているかどうかを判断するため，クラスタの重心の特徴ベクトルとの比較を行う．部分時系列の特徴ベクトルに対して，各クラスタの重心の特徴ベクトルとのユークリッド距離を総当たりで計算する．ユークリッド距離が最も小さくなるクラスタC_kを部分時系列の時系列パターンと定義する．

以上より，テストデータの降水量・気相率の時系列パターンを定義し，開花日を基準として生育段階を区切った後，3.4 節の方法でテストデータの共起パターンデータを作成することができる．

作成したテストデータの共起パターンデータ内に，3.5 節より得られた多収，低収の一方に稀有な頻出アイテム集合が出現するかでテストデータの分類を行う．テストデータ内に現れた多収の頻出アイテム集合の数をTH，低収の頻出アイテム集合の数をTL，分類における閾値をwとするとテストデータの分類ラベルは式 (3.5)で与えられる．

$$\begin{gathered} if\ \frac{TH}{TH+LH} \geq w\ , then\ \text{"high yield"}, \\ else\ if\ \frac{LH}{TH+LH} \geq w, then\ \text{"low yield"}, \\ else\ then\ \text{"uncategorized"} \end{gathered} \tag{3.5}$$

4. 実験

提案手法を実際の降水量データと土壌水分データに適用することで，収量に関与する土壌の排水性や土壌水分環境が発見できるかを検証する．降水量データを導入したことによる有用性を検証するため，入力データとして土壌水分データのみを用いた場合との比較をする．

4.1 実験設定

本実験では，降水量データ，土壌水分データ，実態調査データを用いる．降水量データは，農研機構メッシュ農業気象データシステムが提供する 1 km × 1 km のメッシュ気象データの降水量を用いた．土壌水分データ，実態調査データは，国立研究開発法人農研機構中央農業研究センターより提供されたものであり，全国の圃場からデータを収集している．データの取得日は 2016 年から 2018 年で，実験対象となる圃場は 301 圃場である．実態調査データの精子実重より，精子実重が 250 [kg/10a] 以上を多収圃場，250 [kg/10a] 未満を低収圃場として分類する．301 圃場を分類した結果，多収は 147 圃場，低収は 154 圃場である．降水量，土壌水分の各時系列データは，開花日 1 週間前から開花後 35 日間を使用する．生育ステージの日数は$\Delta t = 7$とし，$r = \{1,2,\dots,6\}$とする．

3.2 節で述べたように，気相率を推定するためには仮比重d_{bulk}と真比重d_{pure}が必要である．しかし，真比重は土壌ごとの変動が小さいため，すべての圃場に対して$d_{\mathrm{pure}} = 2.6$として実験を行った．また，実験圃場では気相率の実測データがないため，本実験では推定データの検証は行っていない．3.3 節におけるサンプリング長 L は降水量，気相率ともに $L = 4$ とした．また，クラスタリングアルゴリズムは k-means 法を使用し，用いる特徴量を表 1 に示す．

表 1：クラスタリングに用いる特徴量

降水量	最大値	最小値	2日目の値 −1日目の値	3日目の値 −2日目の値	4日目の値 −3日目の値	
気相率	最大値	最小値	2日目の値 −1日目の値	3日目の値 −2日目の値	4日目の値 −3日目の値	平均値

なお，気相率においては特徴量間の差が極端に小さいため，平均 0，分散 1 となるように標準化を行った．降水量，気相率のクラスタ数K_p，K_qはそれぞれ$K_p = 10$，$K_q = 11$とした．また，式 (3.4) において，$\alpha = 2$とした．PrefixSpan の最小支持度は 0.03 とした．分類における閾値は$w = 0.75$とした．

4.2 評価指標

提案手法の妥当性を評価するため，性能指標として多収圃場に対する分類精度 $\mathrm{Precision_h}$，低収圃場に対する分類精度 $\mathrm{Precision_l}$ を計算する．そのために 301 圃場のデータに対して 10 分割交差検証を適用する．しかし，本実験では使用できるデータセットが少なく，学習データの選び方によっては実験結果が大きく変わることが予想される．そのため，10 分割交差検証を 100 回行い，それぞれの交差検証での実験結果の算術平均値を本実験の結果とし，評価する．

4.3 実験結果

降水量データを導入したことによる有用性を確認するため，入力データを土壌水分データのみにした場合と比較する．表 2 に実験結果を示す．

表 2：実験結果

	$\mathrm{Precision_h}$	$\mathrm{Precision_l}$
提案手法	0.6631	0.6689
土壌水分データのみ	0.6295	0.6551

4.4 考察

表 2 より，得られた共起パターンによる分類精度は，土壌水分データのみを栽培環境として定義したときよりも良くなった．また，提案手法によって得られた多収，低収要因は降水量の変化に対する土壌水分の変化であるため，土壌の排水性を評価することが可能であり，実際の栽培環境改善の際にも抽出された情報は役立つと考えられる．

5. 結論

本論文では，降水量と土壌水分の時系列データに対して，特徴的，あるいは一般的な時系列パターンを定義し，両時系列パターンの共起に着目することで，収量に関与する土壌水分を得るための手がかりを発見する手法を提案した．

提案手法の有効性を確認するため，実際の降水量データと土壌水分データを用いて実験を行った．提案手法と入力を土壌水分のみにした手法を比較した結果，提案手法の方が多収，低収ともに分類精度が高く，多収圃場に対しては 0.6631，低収圃場に対しては 0.6689 となった．

謝辞

本研究は，農林水産省「収益力向上のための研究開発(中課題番号：15653568，中課題名：多収阻害要因の診断法及び対策技術の開発)」の支援により実施した．

参考文献

[1] K. Higashiyama，R. Nishide，T. Ohkawa，Y. Chonan，S. Hayashi，T. Nakamura，H. Tsuji，N. Murakami，S. Ozawa，“Extraction of Soil Moisture Change Involved in Soybean Yield by Similarity Evaluation Encompassing Time Series Data”，Proceeding of the 7th IIAE International Conference on Intelligent Systems and Image Processing 2019, pp. 278-285 (2019) .

[2] 国立研究開発法人農業・食品産業技術総合研究機構，メッシュ農業気象データシステム，https://amu.rd.naro.go.jp/（最終閲覧日：2021 年 6 月 18 日）．

[3] J. Pei，J. Han，B. Mortazavi-Asl，J Wang，H. Pinto，Q. Chen，U. Dayal，M. Hsu，“Mining sequential patterns by pattern-growth: the PrefixSpan approach”，IEEE Transactions on Knowledge and Data Engineering，pp. 1424-1440 (2004) .