階層ベイズモデルを用いた多変量地盤物性値の適応型欠損補完

Adaptive Missing Data Imputation of Multivariate Geotechnical Properties
with Hierarchical Bayesian Model

斉藤 大雅 *
Taiga SAITO

*先端社会基盤学研究室（指導教員: 京谷 孝史 教授, 研究指導教員：大竹 雄 准教授）

本研究では，地盤工学における地盤パラメータの多変量空間分布推定に焦点を当て，階層ベイズモデル（HBM）の適用可能性とその拡張について検討した．具体的には，Global-BID（CLAY/10/7490）と Local-BID（Tokyo-CLAY/14/67760）という，異なる特性を持つ 2 つの大規模地盤調査データベースを用いて交差検証を実施し，その推定精度を比較評価した．このプロセスにおいて，サイトユニークネスの仮定に基づく階層ベイズモデルの適用に関する課題を特定し，これらの問題に対応するために混合ドメインを考慮した，よりフレキシブルな確率モデルの導入の必要性を見出した．この新しいアプローチは，従来の線形モデルでは捉えきれなかった地盤パラメータ間の高次従属関係を明らかにし，非正規分布，離散分布，カテゴリーデータなどの多様なデータタイプを統合し，より複雑な地盤状況を数理的に表現できる可能性があることを示した．今後の課題としては，普遍的なサイトの再定義，混合ドメインデータの活用，最小情報従属モデルの理論的拡張，大規模言語モデル導入の可能性検討などが挙げられる．これらの課題に対処することで，地盤パラメータ推定の精度向上と地盤工学分野全体の進歩に大きく貢献できる可能性を示した．

Key Words: *Hierarchical Bayesian Model, Soil property prediction, Site identification, Machine learning*

1. はじめに

地盤構造物の挙動は，周囲の土質の性質に大きく依存する．このため，自然地盤が有するサイト固有の不均質性を理解することは，設計や維持管理の応用面において極めて重要な作業である．一般的に，自然地盤のばらつきに対処するため，確率過程理論に基づくモデリングが基本的な考え方であり，地盤の不均質性の評価は地盤信頼性工学における必須な検討事項であると広く認識されている．

この問題に対して，Ching et al., 2021[1] は世界中のさまざまな国や地域から収集された大規模な地盤調査データ（Big Indirect Database; BID）と階層ベイズモデル (Hierarchical Bayesian Model; HBM) を用いて，現場（サイト）で観測される少ない観測データから未観測のパラメータを含む多変量地盤パラメータの空間分布を推定する手法を提案し，実用化されつつある．近年では，BID 内のサイトと対象現場（ターゲットサイト）の類似性を識別するベイズ類似度尺度が Sharma et al., 2022[2] により提案され，従来技術者が経験に基づいて行っていたサイトの「類似性」の識別をデータ駆動的に行うことで，推定計算の計算コストを低減するとともに，推定結果に解釈性を高める研究が行われている．これらの体系的研究は長年課題であった，少量の現場地盤調査から多変量地盤パラメータの空間分布を推定する画期的なアイディアを提供した．著者は，上記フレームワーク（以降，HBM-FW と呼称する）の基本コンセプトに賛同しつつ，より汎化性と地盤工学的解釈性を高めた多変量地盤パラメータの空間分布推定方法の研究開発を目指すことにした．本論文では，HBM-FW の研究で用いられる BID のデータ量的不備に着目する．まず，我が国が有する大規模な地盤調査データベースを構築し，統計的に十分なサンプルを確保し，著者が考える HBM-FW の課題に関する考察を行う．さらに，モデルの科学的進化に基づく高精度化と実用性の向上の観点から今後のモデル高度化のための基礎資料を得ることを目的とする．

2. HBM-FW の基本的な考え方と検討事項

(1) HBM-FW の基礎理論

Ching et al., 2021[1] では，複雑な確率分布形（非正規分布）を有する地盤パラメータを経験的な非線形変換（e.g. Johnson SU distribution）を用いて多変量正規分布空間へ変換している．この空間を本論では理想化空間と呼称する．この理想化空間において，パラメータ同士は線形関係にあり 2 変数間の相関係数により多変量同時確率分布が記述される，という仮定が置かれている．加えて，この空間にはサイトユニークネス性があることを仮定する．すなわち，特定の地域（サイト）内でのデータの類似性は高いが，サイト間のデータは類似性が低いという仮定を設けている．これにより，HBM-FW では，BID 中の任意のサイトで採取されたデータは，それぞれのサイト固有パラメータのガウス分布に従うと仮定しており，さらに，それぞれのサイト固有パラメータはある共通のハイパーパラメータ $\boldsymbol{\Theta}$ に従うと仮定している．そのため，ターゲットサイトデータは，ターゲットサイト固有パラメータに従い，これらは他のサイト内特徴量と同様に，同じハイパーパラメータに従うと仮定している．

数式的に表すと，以下の式 (1) で算出される．

$$p(\mathbf{X}_s^u|\mathbf{X}_s^o,\mathbf{X}_g) = \int p(\mathbf{X}_s^u|\mathbf{X}_s^o,\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\mathbf{X}_g)d\boldsymbol{\Theta} \qquad (1)$$

ここで，$\mathbf{X}_s^u$ と $\mathbf{X}_s^o$ はターゲットサイトにおける未観測量と観測量，$\mathbf{X}_g$ は BID を示す．また，式 (1) の $p(\boldsymbol{\Theta}|\mathbf{X}_g)$ ではハイパーパラメータの推定（HBM-Learning Stage），

$p(\mathbf{X}_s^u|\mathbf{X}_s^o, \boldsymbol{\Theta})$ ではターゲットサイトデータの予測（HBM-Inference Stage）を示す．理想化空間で階層ベイズモデルを構築することにより，階層同時確率分布の条件付き分布は関数で記述することができるため，マルコフ連鎖モンテカルロ法（MCMC）の 1 つであるギブスサンプリングが利用できる．このため，事後分布（推定分布）のサンプリングは，安定かつ効率的に実施することができる．

(2) 検討事項

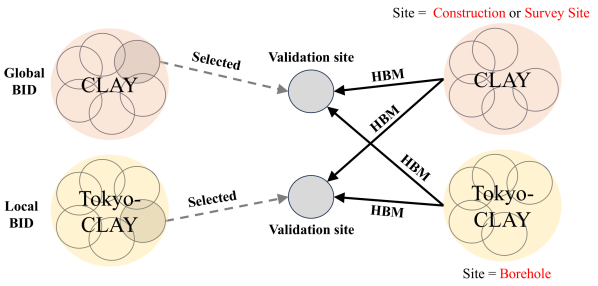
上記の HBM-FW の数理モデルは伝統的な階層ベイズ理論に基づいて完成していると言える．しかしながら，根本的なデータ不足により，下記に示すいくつかの本質的課題が統計学的に検証されていないと考えている．

1) **サイトユニークネスの仮定（サイトの定義）**：理想化空間上にはサイト固有性を有する多変量正規分布が存在し，その集合として BID の多変量地盤パラメータの母集団が形成されているという仮定に基づいている．推定サイトの欠損データは，理想化空間上でサイト間の類似性を評価し，類似性の高い BID のサイトのデータを利用して補完される（Sharma et al., 2022[2]）．ここで，理想化空間への変換は Johnson SU distribution により変換するが，BID 全体のばらつきを正規分布に変換するものであり，個々のサイトが正規分布に従うという仮定を保証するものではない．加えて，BID は世界中の地盤調査サイトの集合体であるが，サイトの定義が曖昧であり，その意味を理解できない点も課題の一つである．
2) **説明変数の選択（モデル選択と解釈性）**：一般化線形モデルなどの一般的な回帰分析では，情報量基準などを利用したモデル選択を行い，回帰式の構造や説明変数の選択を行った最適化なモデルから回帰モデルの意味や統計的考察を行う．一方 HBM-FW は，一般的な回帰分析と異なり，複雑な階層構造を有し，モデル選択の議論の計算コストが高い．また，理想化空間上でパラメータ同士が線形関係になるという前提により，指示的性質（間隙比や含水比など）と力学パラメータ（非排水せん断強度）の関係性を分析することには課題があるように考えている．これらの関係性はより複雑な非線形関係を有していることが考えられ，その部分にサイト固有性が見出せる可能性がある．

3. 研究の方法

図-1 は，本研究の交差検証法（Leave-One-Out）の模式図を示している．世界中の調査データを収録したデータベースである CLAY/10/7490[3] と本研究で整理したデータベースである Tokyo-CLAY/14/67760 を BID として利用し，それぞれの BID から抽出した検証サイトの推定精度を評価する．Tokyo-CLAY/14/67760 は，東京湾に広く堆積する軟弱粘性土から収集した集中サンプリング結果を整理したものである．この BID は，14 種類の地盤パラメータを収録し，1 種類以上の地盤パラメータが観測されている箇所の数が 67760 であり，CLAY/10/7490 より 5 倍以上のデータが存在する．すなわち，前者は世界中の観測データで構成されている Global-BID，後者は平面面積 10km^2 の極めて局所的な範囲からのサンプリングであり，Local-BID と呼称する．

図-2 は，含水比 w と非排水せん断強さ S_u と含水比 w と塑性指数 P_L の散布図を示している．上図は CLAY/10/7490，下図は Tokyo-CLAY/14/67760 の図を示しており，散布図とサイト毎の $\pm 1\sigma$ の範囲を楕円でしてしている．CLAY/10/7490 では，過去に実施されたある調査サイトをサイトの定義として採用し，Tokyo-CLAY/14/67760 では，1 つのボーリングを 1 サイトとして定義している．なお，CLAY/10/7490 のサイト毎の楕円図には Tokyo-CLAY/14/67760 の散布図を薄い着色で重ね書きしている．



図–1 本研究で用いた Leave-One-Out 交差検証法の模式図；それぞれの BID について，任意のサイトを対象に階層ベイズモデルを適用し，推定精度の比較を行う．

Tokyo-CLAY/14/67760 の散布図は，驚くべきことに CLAY/10/7490 の散布図の範囲と概ね一致していることがわかる．世界中の多様な地盤条件の調査データを統合した Global-BID（CLAY/10/7490）とある局所的な高密度地盤調査を収録した Local-BID（Tokyo-CLAY/14/67760）の散らばりがおおよそ一致する点は大変興味深い．ここでは，このように異なる特性を有する BID を利用した HBM-FW を実施し，交差検証結果を比較検討することで先に示した着目点について考察を行う．

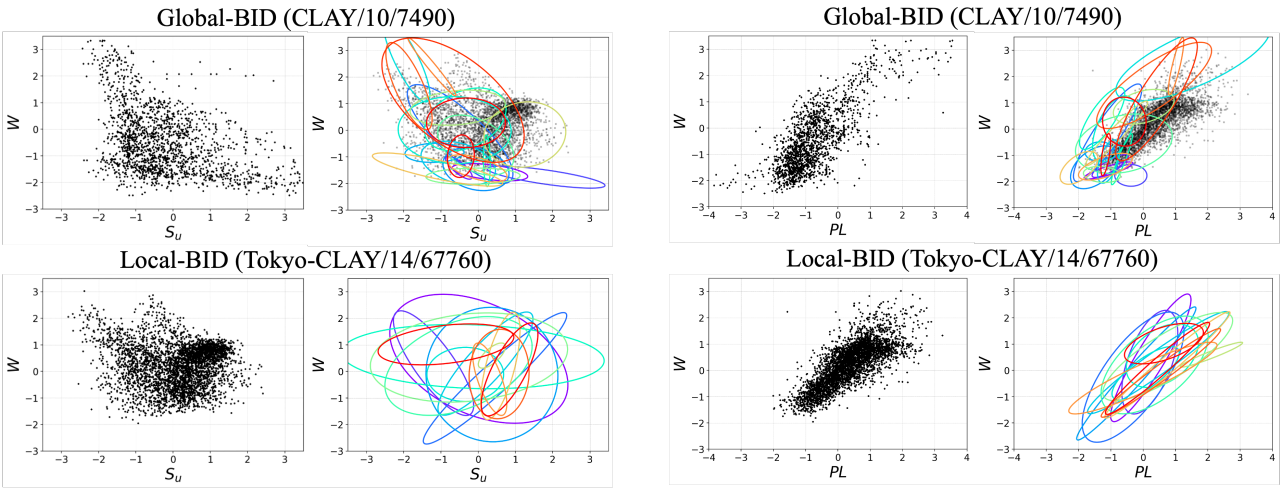
4. 研究結果と議論

(1) 交差検証に基づく考察

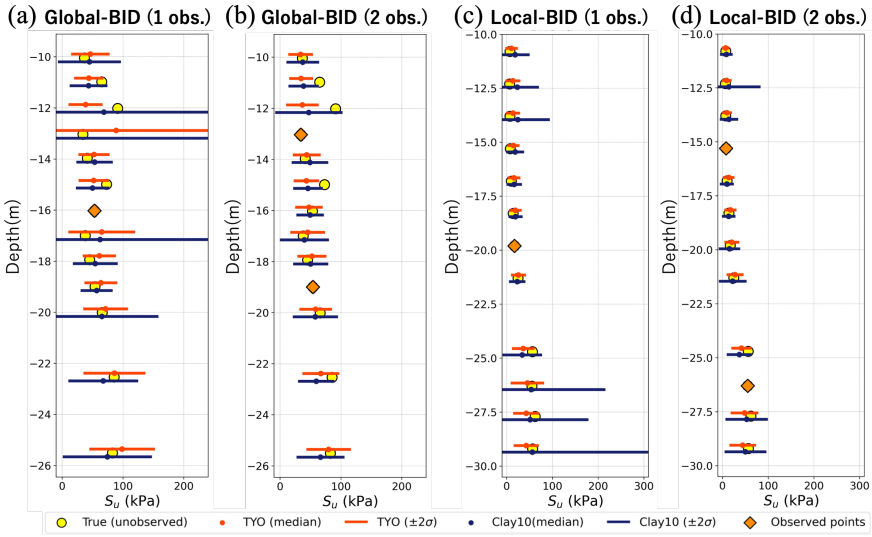
図-2 からわかる通り，CLAY/10/7490 のサイトの定義では，Tokyo-CLAY/10/67760 は 1 つのサイトのデータであると定義できる．ただし，1 つのサイトからの大量サンプリングによるデータの散らばりは，CLAY/10/7490 の散らばりとほぼ同程度である．CLAY/10/7490 の 1 つのサイトに収録されているデータは平均的に 10 程度でデータ数が少ないため，分散が過小に評価されている可能性がある．また，Tokyo-CLAY/10/67760 では，1 つのボーリングを 1 つのサイトと定義してデータを整理している．その確率分布の散らばり（楕円）は 1 つのサイト内でも多様な特性を有していることがわかる．

図-3 は，S_u の深度分布を推定対象として，それぞれの BID から選択した代表 1 サイトの深度分布の交差検証結果を示している．推定結果は中央値と $\pm 2\sigma$ の幅で示しているが，いずれの場合も正解値を概ね適切に捉えており，BID の違いは推定精度に大きな影響を及ぼしていないことが分かる．

Tokyo-CLAY/14/67760 のような Local-BID でもある程度の多様性を有するデータベースの階層構造が構築でき，地盤パラメータの推定にある程度機能するようである．一方で，Sharma et al., 2022[2] が提案する類似性によるサイト識別の導入による，推定の解釈性の向上を考えた時，現場曖昧なサイトの定義を明確にする必要がある．加えて，$\pm 2\sigma$ の幅が局所的に極めて大きい場合が



図–2 2 つの BID の比較プロット；2 つの BID に共通する代表物性値の散布図とサイト毎のガウス密度を示している．



図–3 2 つの BID からそれぞれ選ばれた代表 2 サイトにおける，S_u 深度分布推定結果；(a) は S_u を 1 点観測，(b) は S_u を 2 点観測した時の結果を示す．また黄色プロットは真値，橙色菱形プロットは観測点を示す．階層ベイズ推定の予測結果は，青色（Local-BID を利用した場合）と赤色（Glocal-BID を利用した場合）の横線で示され，中央値は丸点で，その長さは $\pm 2\sigma$ 区間を示している．

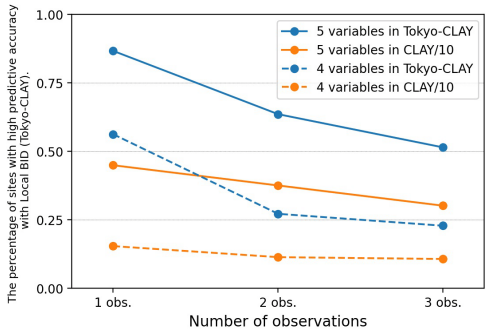
ある．これは主に BID を CLAY/10/7490 とする場合に生じるケースが多い．これは，Johnson SU distribution への変換の影響が懸念される．データ数が多い Tokyo-CLAY/14/67760 を重視した変換が採用されているため，CLAY/10/7490 の一部地点では不適切な分散が評価されている可能性が指摘できる．

図-4 は，深さを説明変数として利用する場合としない場合の交差検証の推定精度をまとめた図である．ここでは，CLAY/10/7490 と Tokyo-CLAY/14/67760 の両方の BID から S_u を推定する問題に対して，深さ *Depth*，含水比 w，塑性指数 P_L，液性限界 LL を説明変数とする場合（5-variables）と深さ *Depth* を説明変数から除いた場合（4-variables）に対して，観測点数と平均尤度（推定精度）の変化を見た図である．いずれの場合も，深さ *Depth* を説明変数にする場合としない場合で推定精度が大きく異なり，推定精度は説明変数に強く依存することが分かる．ここでは，深度依存性が顕著な非排水せん断強度 S_u の推定問題を対象としており，深さ *Depth* が強い説明性を有していることは自明であるが，この結果はモデル選択の重要性を示唆していると考えている．

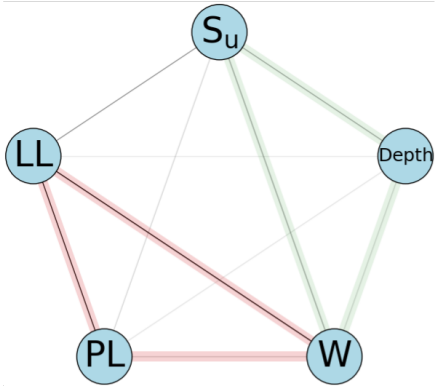
(2) 推定モデルの理論拡張へ向けての議論

上記の大規模な交差検証により，HBM-FW の基本的な有用性を確認した．その上で，BID と統計的機械学習を融合した多変量地盤パラメータ推定問題をより科学的に高度化する観点で，下記の洞察を得られた．

1) **普遍的なサイトの定義**：世界中の様々な調査現場からデータベースを構築する必要があるため，サイトの定義は曖昧である．このことが，サイト識別の効果を低減させ，また，曖昧なサイトの定義はサイト識別結果の解釈性も低減させてしまう可能性がある．基本的にはデータに即したクラスタリングにより，地盤種別や地盤の形成メカニズムに関連したサイトを再定義することが重要である．
2) **根本的な情報不足（混合ドメイン）**：上記の課題を解決する際，現状の連続量ベースの BID 構築には限界がある．地盤の種別や形成メカニズムを反映した地盤パラメータ推定において，土質分類な



図–4 HBM を用いた 2 つの BID の交差検証；横軸がターゲットサイトにおける S_u 観測点数，縦軸が Local-BID サイトによる階層ベイズ推定が優位になるサイトの数の割合を示す．青実線が Local-BID を対象とした 5 変数の階層ベイズ推定結果，橙実線が Global-BID を対象とした 5 変数の階層ベイズ推定結果を示す．またそれぞれの破線は 4 変数の階層ベイズ推定結果を示す．



図–5 地盤物性値間の高次従属ネットワーク図；2 変数間のエッジ（黒線）の太さは従属性の強さを表す．3 変数間で最も従属性が高いハイパーエッジを赤，次に従属性が高いハイパーエッジを緑で示す．

どのカテゴリーデータ，土のコア画像，弾性波探査などの物理探査結果を含む高次元空間データなど多様なデータタイプを統計モデルに反映させること（混合ドメイン解析）が重要である．

上記を踏まえて，著者は，混合ドメインを許容するより汎用的な確率分布の地盤工学への応用についての基礎研究に着手している．詳細は割愛するが，まず，最小情報量従属モデルという新しい数学概念の導入を試みる．最小情報従属モデルの基本式は下記の通りである．

$$p(x;\theta,\nu) = \exp\left(\theta^{\top}h(x) - \sum_{i=1}^{d} a_i(x_i;\theta,\nu) - \psi(\theta,\nu)\right)\prod_{i=1}^{d} r_i(x_i;\nu) \tag{2}$$

d を物性値数（次元）とした時，$r_1(x_1;\nu_1),...,r_d(x_d;\nu_d)$ は，$1,...,d$ における変数それぞれの周辺密度の統計モデルであり，$\nu_1,...,\nu_d$ はそれぞれの周辺密度を特徴づけるパラメータを表す．また $\theta \in \mathbb{R}^K$ は従属性を表す K 次元パラメータ，$h : \mathcal{X} \to \mathbb{R}^K$ は所与の関数を表す．これらは，条件付き尤度を用いた MCMC 交換アルゴリズムを適用することで，陽に ν を同定する必要なしに擬似的に θ を算出することができる．これにより，任意の変数間にある従属性の把握が期待される．また，関数 $a_i(x_i;\theta,\nu)$ と $\psi(\theta,\nu)$ は確率分布の制約を満足するように定めることができる．非正規分布，離散分布，カリゴリーデータを統一的に取り扱うことができるため，これまで利用してこなかった地盤情報を利用したパラメータ推定へ拡張が期待される．また，理想化空間への変換を必要としないことから，地盤情報が有する複雑性をそのままモデル化し，サイト識別をより高度に実施できる可能性がある．

また，このモデルは，2 変数以上の複数のパラメータ間の従属関係を分析することができる．**図-5** は，先に示した 5 つの地盤パラメータの高次依存関係を分析した結果である．非排水せん断強度 S_u が強い深さ依存性を有していること，塑性指数や液性指数には従属関係がないことが確認できる．このグラフ自体が従来の線形ベースの統計モデルでは描くことができない情報であり，サイト固有性を示す重要な情報になると考えられる．この事例は，地盤工学的に自明の関係を分析しているが，関係性が不明な多様な混合ドメインデータの解析へ拡張することにより，サイトの固有性をより高度にモデル化することができる可能性がある．

5. 結論と今後の課題

本研究では，階層ベイズモデルを用いた地盤パラメータの多変量空間分布を推定する手法（HBM-FW）を詳細に考察し，今後のモデル化の高度化の可能性を研究した．Global-BID（CLAY/10/7490）と Local-BID（Tokyo-CLAY/14/67760）を用いた交差検証により，異なる特性を持つデータベース間での推定精度に大きな差異がないことを確認し，HBM の有用性を示した．また，その一方で，サイトユニークネスの仮定の下でのモデルの適用とその解釈性に関する課題を明らかにし，これらの問題への対応として，混合ドメインデータを取り扱える最小情報従属モデルの導入を試みた．この新しいアプローチは，従来の線形モデルでは捉えられなかった地盤パラメータ間の高次従属関係を明らかにする可能性を示唆している．

今後の研究においては，土質分類，物理探査結果などの多様なデータタイプを統合することで，地盤パラメータ推定の精度向上と解釈性の向上を目指す．このために，カテゴリーデータや離散データを含む高次元データの分析手法を開発する予定である．さらに，最小情報従属モデルの適用性の検討と，その理論的基盤の強化，応用範囲の拡大を目指す．また，大規模言語モデルの埋め込み技術を利用した言語の数値変換に通じて，対象サイトの地形地質の形成過程やサイト固有の性質を統計解析に取り入れ，因果関係の明確化を通じて，地盤工学における新たな知見の獲得に務める．

参考文献

1) Jianye Ching, Stephen Wu, and Kok-Kwang Phoon. Constructing quasi-site-specific multivariate probability distribution using hierarchical bayesian model. *Journal of Engineering Mechanics*, Vol. 147, p. 04021069, 10 2021.
2) Atma Sharma, Jianye Ching, and Kok-Kwang Phoon. A hierarchical bayesian similarity measure for geotechnical site retrieval. *Journal of Engineering Mechanics*, Vol. 148, p. 04022062, 05 2022.
3) Jianye Ching and Kok-Kwang Phoon. Transformations and correlations among some clay parameters — the global database. *Canadian Geotechnical Journal*, Vol. 51, No. 6, pp. 663–685, 2014.

(2024 年 1 月 30 日 提出)