データ規模とモデル選択に関する一考察: 区間別Fundamental Diagram推定を例に

中西 航 [1]

[1]正会員　金沢大学准教授　理工研究域　地球社会基盤学系 （〒 920-1192 石川県金沢市角間町）
E-mail: nakanishi@se.kanazawa-u.ac.jp (Corresponding Author)

土木計画学の様々な分野において，大規模で高解像度なデータの利用可能性が高まり続けている．このとき，統計学で一般に用いられるモデル選択規準は，サンプル数が増えるにつれて変数の多い複雑なモデルを支持する傾向がある．しかし，これは本分野がモデル推定に期待する問題解決の方向性と整合しないように思われる場面もある．すなわち，我々が直面する具体的な問題において，データ規模に応じたモデルの複雑さをどう捉えるべきか，その際にどのような課題が存在するかは必ずしも整理されていない．そこで本稿では，著者既報であるマクロ交通流の Fundamental Diagram を地点ごとに推定する問題を例に，データに関連する課題を整理し，統計的モデル選択規準の現状を説明し，規準の性質を実例で示したうえで，今後の展望を述べる．

Key Words : *sample size, model selection, information criterion, regularity and singularity, location dependent fundamental diagram*

1. はじめに

(1) 交通流の分析と大規模・高解像度データ

土木の多くの分野に違わず，交通流の観測やモデリングにおいても大規模で高解像度なデータの利活用が進んでいる．たとえば，多数の地点でのトラフィックカウンターの 5 分交通量データが提供されている [注1),注2)]．これらは，研究はもちろんのこと，各種コンペティション等でも用いられている [1), 注3)]．

本稿は，そのようなデータの一例として，全車両軌跡データを扱う．一般に，車両軌跡データとは，コネクティッドカー[1]の位置を時々刻々と記録したものである．その記録間隔が十分に小さければ，実質的に当該車両の走行軌跡を完全に把握可能であるといえる．全車両軌跡データとは，文字通りすべての車両からこの情報が得られる場合のデータのことである．あらゆる空間において全車両軌跡データが得られれば，少なくとも[2]交通工学の研究上は理想的だが，この想定は現実的ではない．現時点では，特定の道路リンク上の特定の時間帯に限ってそのようなデータが得られている．具体的には，NGSIM[2)] は約 600 m×15 分間 ×6 セット，Zen Traffic Data[3), 注4)] は 1.7-2.8 km×1 時間 ×16 セット，I-24 MOTION[4)] は 6.75 km× 日中全時間帯という規模感である．日本では，阪神高速道路株式会社が提供する Zen Traffic Data が最大のデータセットであり，既存モデルの大規模な検証や新規モデルの開発，シミュレーションモデルへの応用等に利用されている [5),6),7),8),9),10),11),12),13)]．

(2) Fundamental Diagram の高解像度化

上記のような大規模で高解像度なデータを用いて，マクロ交通流分析の根幹をなす流率-密度-速度関係 (Fundamental Diagram, FD) を空間的に高解像度に推定する方法を検討する．道路の性能を，その道路上で実現する交通状態 (流率 q，密度 k，速度 v の 3 変数の組) の関係性で記述するとき，FD は定常状態 (**2.(2)b)**) にある交通状態が経験的に従う関係性である: 密度 k を引数とする関数 f を用いて $q = f(k)$ や $v = f(k)$ のように表現したものである．FD は道路ネットワーク上の交通量の分布や渋滞の伝播などを把握・予測するために大きな役割を果たす [14),15)]．ゆえに，実データに基づいて高精度な FD を推定することは工学的に重要である．

FD は交通状態の 3 変数のみならず，地点 x，時刻 t，個人 n にも依存する (i.e., $v = f(k, x, t, n)$) [16),17)]．観測データに基づく FD 推定では，分析者がこれらの変数について暗黙裏に仮定を置く．たとえば，トラフィックカウンタの設置地点 $x = X$ で長期間観測したデータを FD 推定に用いる場合，地点 $x = X$ の FD を時刻 t や個人 n に依存しない単一のものとみなして求めることに相当する．

さて，従来はデータを観測できる地点が限られていたため，地点 $x = X$ で推定した FD を，その周辺に共

[1] 日本ではプローブカーと呼ばれることが多い．
[2] たとえばプライバシーの問題は別に存在する．

通の FD として扱ってきた．また，観測ができない地点における交通状態の推定は重要な課題であり，多様な切り口から多くの研究が蓄積されてきた[18),19)]．一方で，すでに述べたとおり，近年は全車両の軌跡データのような時空間的に高解像度なデータが入手可能となった．このデータを利用すれば，原理的には任意の地点における FD を推定できる．たとえば，直線・曲線区間，任意の微小区間ごとの FD である．以降，この FD を区間別 FD と呼ぶ．FD がネットワーク上の分析に用いられることを考えれば，現象の正確な把握や予測にあたって多くの地点で正確な FD を知ることが望ましい．その際に区間別 FD の考え方は有用であろう．

しかし，車両軌跡データが与えられた連続空間に対して，相互に異なる FD を推定すべき区間を適切に設定することは難しい．区間を不必要に小さく設定すると，本来は同一の FD に従うひとつの区間が異なる区間に分離されてしまう．すると，観測データに内包されるバラツキにより，分離された区間どうしに本来有意ではない微小な差異を見いだすことになる．これは統計的に不適切なばかりか，推定された区間別 FD を用いたその後の分析を複雑化する点で実用性を欠く．一方で，区間を大きく取りすぎれば，実際には異なる FD に従っている区間を同一の FD に従う区間として集約することになる．これは現象理解として正しくなく，FD を利用した分析結果も不正確になるばかりか，高解像度データが有する豊富な情報を無駄にすることにもなる．

(3)　本稿の主題と構成

この問題の解決方法は，全車両軌跡データが与えられた際に適切な区間分割を行う方法を構築することである．そして，線形な FD 関数の仮定下では，相互に異なる FD に従う区間を特定し，同時にそれらの区間別 FD を推定する手法が提案されている[20)]．これにより，分析者の先験的な知識を援用しなくても統計的に最適な区間別 FD を推定できる．また，実用性の高い任意の区間数で推定することも可能な手法である．

ここで，この例を少し抽象化して捉えたい．すると，「大規模なデータが与えられたときに，どのくらい・どのように複雑なモデルが適切か」という問題を考えていることが分かる．「どのくらい」が区間数に，「どのように」が区間の分かれ方にそれぞれ相当する．また，区間数が多いほどパラメータ数が増え，「複雑なモデル」になることも直感的に理解できよう (具体的には **4.(3)**)．

さて，**5.(4)** で説明するように，一般にデータが大規模になると複雑なモデルが支持されやすくなる．目の前で起きている現象は何も変わっていないにも関わらず，観測が大規模になるだけで最適とされるモデルが複雑になっていくということである．区間別 FD 推定においては，最適な区間数が増え，区間の分かれ方が変化しうることになる．そして，このことは区間別 FD 推定に限った話ではない．むしろ，大規模データを扱うすべての問題に共通する事象であり，土木計画学の多くの分野において重要な論点であると考えられる．

そこで本稿は，区間別 FD の推定問題を通して，データが大規模になると複雑なモデルが支持されるということについて考えていきたい．そのために，まず **2.** で大規模で高解像度なデータを用いた FD 推定および周辺の課題を整理する．続く **3.** で全車両軌跡データと FD 推定に用いるサンプル生成を説明し，**4.** でスパースモデリングを用いて区間別 FD をモデル化する．その後，実際の推定に先立って **5.** では統計的なモデル選択の現状を概観する．そして，**6.** でシミュレーションデータ，**7.** で実データに対してそれぞれ適用および考察を行う．**8.** では今後の展望を含めつつ本稿をまとめる．

2.　課題の整理

本章では，大規模で高解像度なデータを用いた FD 推定および周辺の課題を整理する．もちろん，以下に示すもの以外にも，多種多様な課題が存在すると思われる．著者の能力不足や紙面の制約によりそのすべてを扱うことはできないため，本稿 **5** 章では特に **(5)** のモデル選択を扱う．

(1)　データ取得

存在するデータが限られている場合，そこから分析者が研究に用いるデータを決定することは容易である．一方，すでに述べたように，大規模で高解像度なデータが多数蓄積されていることは確かである．しかしながら，その全容，すなわちどのようなデータがどの組織に存在するのかの把握は容易ではない．また，日本では，データのオープン化が海外の同分野あるいは国内の他分野に比べて進んでいない状況である．たとえば，データの公開期間が限定的である，データの利用申請が必要である，コンペティションに参加した場合のみ用いることができるなどの制約が存在することは珍しくない．そのため，分析に適したデータを探し，利用可能とする過程自体に検討が必要となっている．

(2)　データ処理

全車両軌跡データ等から交通状態を算出し，FD 推定のためのサンプルに加工する処理に相当する．

a)　サンプルデータ生成

従前のトラフィックカウンターのデータを用いる場合を考えよう．観測地点は当該カウンターの地点に限定される．また，観測できる交通状態変数は流率と (密

度の代替変数である) 時間占有率であり，機器によっては車両個々の速度も含まれる，などということが事前に明らかであった．これに対し，いわば生データである全車両軌跡データを用いれば，任意地点・時点の交通状態変数を算出できる．そのため，分析に用いる際には分析者が何かしらのルールを定め，データ処理を行う必要がある．

たとえば，任意の地点で仮想的なトラフィックカウンターを設けたものとしてデータを生成することができる．あるいは，任意の時空間領域において Edie の定義[21] に基づいて交通状態変数を算出することもできる．後者は一般性の高い手法であり，理論的に妥当であろう．ただし，その時空間領域の形状や大きさを決定する必要がある．領域の形状を矩形に限定したとしても，時間・空間方向それぞれの大きさを決めなければならない．この大きさの設定方法に普遍的な指針は存在しない．そこで，多様な設定方法による FD 推定の検討がなされている[10]．なお，本稿が扱う区間別 FD では，空間方向の大きさが区間の最小長さとして扱われることになる (詳細は **3.(2)**)．

b)　データクリーニング

一般にデータからモデルを推定する際に，異常値等を取り除くデータクリーニングが行われる．FD 推定においては，定常判定という操作がこれに対応する．FD は直下で述べる定常状態にある交通状態が従う関数である．したがって，前項で算出した交通状態から定常状態にあるもののみを抽出する必要がある．これを定常判定という．定常状態とは，対象時空間領域内の全車両が等速度かつ等車頭距離で走行している状態である．ただし，厳密な意味でそのような交通状態が観測されることは稀である．そこで，多くの研究では「ほぼ定常 (near steady)」でこれを代用し，速度の変動係数を用いて定常判定を行っている[10),22),23)]．その具体的な手順は **3.(3)** で詳述する．

ただし，この方法は，必ずしも全車両軌跡データのような高解像度データを念頭に置いたものではない．現に，この定常判定を行った後にも大きなデータのばらつきが残存することがある (たとえば後出の図–**3**)．特に前項の矩形を小さく設定した場合 (i.e., より高い解像度で分析を行おうとした場合) に顕著な傾向である[注5)]．この理由のひとつは，FD の個人依存が考慮できていない点や等車頭距離が考慮されていない点にある．定常判定方法の改善に取り組んでいる研究も存在する[注6)]．

(3)　回帰モデルの設定

a)　FD 関数

1935 年の Greenshields[24] 以来，多種多様な FD 関数が提案されてきた．理論的な解釈可能性，データへの当てはまり，パラメータ推定の容易さ，あるいはネットワーク分析への展開可能性などの観点でそれぞれ利点や欠点を有する．また，前述の通り，交通状態変数の算出方法と FD 関数の当てはまりの良さが相互に関係することも分かっている[10]．

b)　被説明変数

仮に観測データが流率 q と速度 v であるとしよう．ここで，FD 関数が $v = f(k)$ の形式で記述されているならば，その推定に先立って恒等式 $q = kv$ を用いてデータを k と v の組に変換する必要がある[3]．この場合，結局は $v = f(q/v)$ という式に対してパラメータ推定を行っていることになる．もちろん推定は実行可能だが，両辺に同じ観測変数が投入されることへの違和感は否めない．また，被説明変数に q と v のどちらを設定するかにより，最適化の目的関数も異なる[4]．しかしながら，観測データと被説明変数の取り方の関係について詳細に論じられることは少ないように思われる．その主たる理由は，実務上の問題が少なく，観測データ制約が大きければ議論の余地も少ないことだろう．これは高解像度データゆえに追求が可能になった課題である．

c)　追加的な説明変数

データの規模や解像度によっては，FD 関数の引数に個人 n や時刻 t を導入することも理屈上は可能である．ただし，時間帯ごとの違いや個人差に由来するばらつきについて個別の議論はある[8),25)] ものの，包括的な検討は著者の知る限り行われていない．

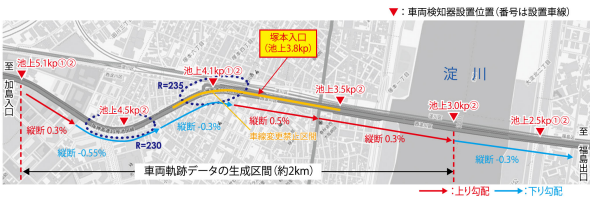
(4)　パラメータ推定

前節で少し触れたが，FD 関数にはパラメータが含まれており，観測データから推定する必要がある．最小二乗法や最尤推定法を用いることができる場合には，それらを用いることが多いようである．ただし，誤差分布に (明示的か否かを問わず) 正規分布を仮定するため，高密度領域では現実には発生し得ない負の流率や速度を許容する状態になっている．このことが FD の精度にどの程度・どのような影響を及ぼしているかは定かでない．また，自由流領域と渋滞流領域で異なる関数に従う FD[5]の場合，必然的に段階推定，EM アルゴリズムあるいはベイズ推定の導入が必要となる．

(5)　モデル選択

モデル選択とは，データへのモデルの当てはまりの良さを比較検討する作業である．この「当てはまりの良さ」の基準が複数あるために，モデル選択に用いら

[3] 交通状態の 3 変数の間には，常に $q = kv$ が成立する．
[4] たとえば，最小二乗法であれば q と v のどちらの残差二乗和を最小化するのかが異なる．当然の帰結として，推定される FD も若干異なるものとなる．
[5] たとえば，頻繁に用いられる Triangular FD も含まれる．



図–1 対象区間 (資料 注4) より引用)

れる規準もまた複数存在する．この節では概略を述べ，詳細は **5** 章で議論する．

複数の FD 関数候補を想定した場合，それらのなかから当てはまりの良いものを選ぶ方法が必要である．残差二乗和や尤度に着目する方法，各変数の有意性に着目する方法，モデル全体の性能に着目する方法などが存在する．また，変数の候補が大量に存在する場合，その全パターンを試すことは組合せ爆発により非現実的である．そのため，LASSO[26] に代表されるスパースモデリングを用いることもある．本稿で扱う区間別 FD もこれに相当する．

(6) 精度検証

前節は，あくまでも統計的なモデルの良さを述べたものである．一方，実用面からは，FD 推定による交通流の再現性もまた重要な観点である．得られた (区間別)FD を用いたシミュレーションにより軌跡や交通状態がどの程度再現できるかを検証することが考えられる．また，前出の未観測地点における交通状態推定手法についても，実データによる検証が可能となる．これにより，実用的な意味で必要な観測の解像度を議論することができる．一定程度の観測データにより未観測地点を十分精度良く推定できることがわかれば，実務上大きな価値があろう．

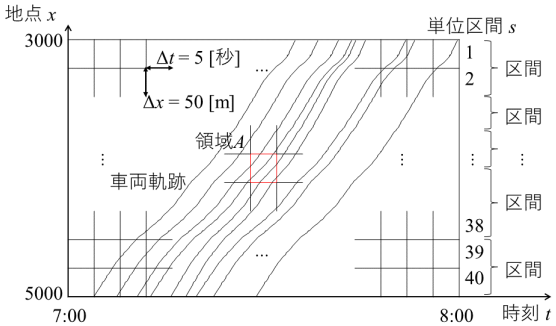
3. 使用データ

本章では，全車両軌跡データから以降の FD 推定でサンプルとして用いる交通状態を算出する．基本的に既報[20] と同様であるため，ここでは概略を示す．

阪神高速道路株式会社が提供する全車両軌跡データ Zen Traffic Data のうち，11 号池田線の大阪方面 (上り線) の塚本合流付近における，平日午前 7-8 時の追越車線を対象とする．5.0 キロポストから 3.0 キロポストの 2 km にわたって全車両の走行位置が記録されている．1745 台が通過しており，出勤ラッシュに起因すると思われる渋滞状態も観測されている[8](**図–1**)．

(1) 全車両軌跡データ

以降では具体の場所を指し示すために，3.0 キロポストの地点を $x = 3000$ [m]，3.5 キロポストの地点を



図–2 交通状態の算出と単位区間の設定

$x = 3500$ [m] のように，変数 x を用いてメートル単位で表記する．先験的には，上り・下り勾配や勾配変化に伴うサグ部，入口車線の合流部，車線変更禁止区間などが FD に影響を与えうる．

(2) 交通状態変数の算出

交通状態の 3 変数 (流率 q，密度 k，速度 v) を Edie の定義[21] に基づき算出する (**図–2**)．具体的には以下:

1. 時空間上に時間幅 Δt および距離 Δx で囲まれる矩形領域 A を定義する．
2. 領域 A 内の全車両の総移動距離と総移動時間を集計する．
3. 以下の式で交通状態を定義する:

$$q = \frac{\text{総移動距離}}{|A|} \tag{1}$$

$$k = \frac{\text{総移動時間}}{|A|} \tag{2}$$

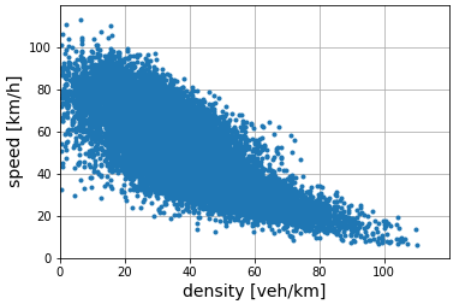
$$v = \frac{\text{総移動距離}}{\text{総移動時間}} \tag{3}$$

ただし，$|A| = |\Delta t \Delta x|$ は領域 A の面積である．

領域 A は $\Delta t = 5$ [秒] および $\Delta x = 50$ [m] で構成する． そして，データの全時空間領域 (1 時間 $\times$ 2 km) を埋め尽くすように $720 \times 40 = 28800$ 個の領域に分割し，それぞれについて交通状態を算出する．以降では距離 Δx ごとに区切られた部分を単位区間と呼び，下流から順に $s = 1, 2, \ldots, 40$ と番号を振る．すなわち，単位区間 s は地点にして $2950 + 50s \leq x < 3000 + 50s$ [m] の範囲を意味する．また，連続する単位区間 1 つ以上からなる部分を区間と呼び，その区間における単一の FD を区間別 FD とする (**図–2**)．すなわち，単位区間の長さは，区間別 FD が取りうる最小の長さとなる．

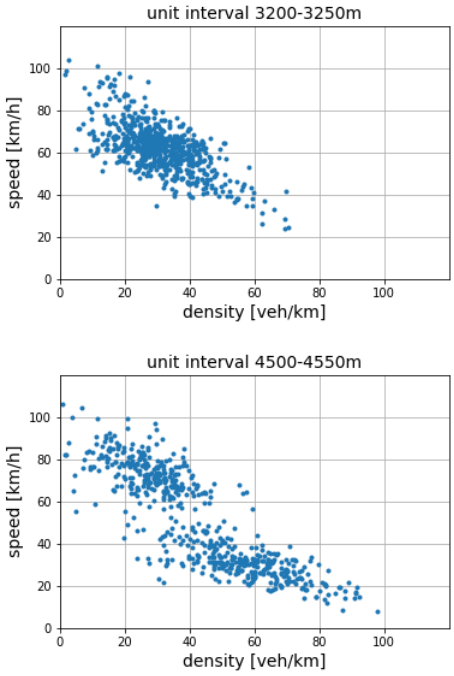
(3) 定常判定

前述の定常状態を，全車両が等速度という条件に着目し，以下の方法で判定する．



横軸が密度 k [台/km]，縦軸が速度 v [km/時].
以降の図も同様.

図–3 定常状態での速度-密度関係



上図: $s = 5$，サンプル 707 個.
下図: $s = 31$，サンプル 561 個.

図–4 単位区間における定常状態での速度-密度関係の例

1. 領域 A 内に存在する車両それぞれについて，個別に速度 (=領域 A 内の移動距離/領域 A 内の移動時間) を算出する.
2. これを用いて, 領域 A 内に存在する車両の速度の変動係数 (=車両どうしの散らばりを表す標準偏差/車両間の平均) を求める.
3. 変動係数が事前に定めた閾値 $\theta(= 0.15)$ 以下の交通状態を定常状態と判定する.

得られた定常状態を $k-v$ 関係の散布図として**図–3** に示す．定常状態と判定された交通状態は 25150 個であった．単位区間別の $k-v$ 関係の一例として，$s = 5$ および 31 のデータのみの散布図を示す (**図–4**)．両区間の散布図を視覚的に比較すると，同一の FD に従うとみなすよりは，それぞれに異なる FD を推定したほうが良いようにも思われる．次章のモデル化によりこのことを統計的に扱い，最適な区間設定手法を構築する.

4. 区間別 FD のモデル化

(1) 前提条件

推定する区間別 FD について，以下の仮定を置く:

1. FD の関数形は，簡単のため Greenshields の式[24] を用いる：速度 v が密度 k の一次関数として
$$v = v_f + wk \tag{4}$$
である．未知パラメータは v_f と w の 2 つで，それぞれ切片 ($k = 0$ のときの v の値) と傾きを表す.
2. 区間によらず，FD の関数形は同一とする．すなわち，異なる区間別 FD どうしでは，上式における切片と傾きの値のみが異なる.
3. 隣接する単位区間で切片や傾きの値が大きく異なるようなデータのとき，それらは異なる FD に従う区間に所属する．そうでない場合は，両単位区間は同一の FD に従う区間に所属する.

(2) Generalized Fused Lasso

区間別 FD を，スパースモデリング手法の一種である Generalized Fused Lasso[27] を用いて定式化する．これは空間統計分野での適用でいくつかみられる[28),29)] 手法である．スパースモデリングの目的は，回帰モデル特定に際する適切な説明変数の選択 (およびその結果としての過学習の回避) である．そのために，大量の説明変数の候補から少数の重要な説明変数を特定する手続きを制約付き最適化問題として定式化する.

スパースモデリングの代表的手法に Lasso[26] がある．Lasso では，回帰モデルの損失関数とパラメータの L_1 正則化項との重み付き線形和を最小化するパラメータを求めることで，変数選択とパラメータ推定とを同時に実行する．具体的には，モデル式を
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5}$$
とするとき, 以下の式を最小化するパラメータを求める.
$$\min_{\boldsymbol{\beta}} \left[\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right] \tag{6}$$
ただし，$\boldsymbol{y}$ は N 行 1 列の被説明変数ベクトル，$\boldsymbol{X}$ は N 行 K 列の説明変数行列，$\boldsymbol{\beta}$ は K 行 1 列のパラメータベクトル，$\boldsymbol{\varepsilon}$ は N 行 1 列の誤差項ベクトルである．$\|\cdot\|_1, \|\cdot\|_2$ はそれぞれ L_1, L_2 ノルムを表し，λ は回帰モデルの損失関数に対するパラメータの L_1 正則化項の重みを定めるハイパーパラメータである．正則化項の性質により，重要でない変数のパラメータは 0 に推定されやすい．このとき，その変数候補は選択されなかっ

たと判断する．$\lambda = 0$ のときは，すべての変数候補を採用する回帰モデルに帰着する．λ がきわめて大きいときは，損失関数の L_1 正則化項が卓越し変数は採用されなくなる．なお，正則化項の影響を各変数に対して均等にするために，説明変数を投入する際には平均 0，分散 1 に標準化する．

この正則化項を工夫することで，Lasso の様々な拡張形が提案されてきた．Generalized Fused Lasso は，隣接するパラメータの差についての L_1 正則化項を式 (6) に加える．パラメータの隣接関係は分析者が任意に定義可能である．そして，以下の式を最小化するパラメータを求める．

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{K} \beta_j X_{ij} \right)^2 + \lambda \sum_{(\mu,\nu)\in C} |\beta_\mu - \beta_\nu| + \gamma\lambda \sum_{j=1}^{K} |\beta_j| \right] \tag{7}$$

ただし，y_i は $\boldsymbol{y}$ の i 行目の要素，X_{ij} は $\boldsymbol{X}$ の i 行 j 列の要素，β_j は $\boldsymbol{\beta}$ の j 行目の要素，C は隣接関係にある組み合わせの集合である．γ はパラメータの差の正則化項に対するパラメータ自身の正則化項の重みを定めるハイパーパラメータである．2 つの正則化項の働きにより，重要でない変数のパラメータが 0 に推定されやすいだけでなく，隣接するパラメータの差も 0 に推定されやすい．

(3) 定式化

先に示した前提条件の 1. と 2. により，定常な交通状態 (v_i, k_i) は以下で回帰できる:

$$v_i = \sum_{s\in S} I_{is} \left(v_{fs} + w_s k_i \right) + \varepsilon_i \tag{8}$$

ただし，S は単位区間の集合，I_{is} は交通状態 i が単位区間 s で得られたものである場合に 1(それ以外の場合に 0) となるダミー変数，v_{fs} と w_s はそれぞれ単位区間 s における切片と傾き，ε_i は誤差である．

7 章の実データへの適用では，上式で 40 個の単位区間それぞれに固有に与えた v_{fs} と w_s の計 80 個を変数候補のパラメータと考える．そのうえで，前提条件の 3. を念頭に，隣接する単位区間の同種のパラメータを式 (7) における隣接するパラメータとみなす．C の要素を $(\mu,\nu)=(s', s'+1), s' = 1, 2, \ldots, 39$ としたときに，式 (8) を式 (7) に代入することで，本研究で求めるパラメータは以下の式を最小化するものとなる：

$$\min_{\boldsymbol{v_f},\boldsymbol{w}} \left[\frac{1}{2} \sum_{i=1}^{N} \left(v_i - \sum_{s\in S} I_{is} \left(v_{fs} + w_s k_i \right) \right)^2 + \lambda \sum_{(\mu,\nu)\in C} \left(|v_{f\mu} - v_{f\nu}| + |w_\mu - w_\nu| \right) + \gamma\lambda \sum_{s\in S} \left(|v_{fs}| + |w_s| \right) \right] \tag{9}$$

ただし，$\boldsymbol{v_f}, \boldsymbol{w}$ はそれぞれ v_{fs}, w_s を並べたパラメータベクトルである．式 (9) における γ は v_{fs} や w_s を 0 に推定されやすくする度合いを表す．しかし，本稿の範囲ではある単位区間 s における切片や傾きを 0 に推定したい状況は実際には起こりにくいため，$\gamma = 0$ に固定する[6]．より複雑な FD の関数形であれば，特定のパラメータ自身を 0 と推定したい場合もあるだろう．

さて，式 (9) の第 2 項の性質から，隣接する単位区間の v_{fs} や w_s は同一の値に推定されやすい．いくつかの連続する単位区間で同一の値が推定されるとき，それらは同一の区間別 FD に従うことを意味する．これにより，同一の FD に従う区間を設定できる．なお，このような正則化項を用いる定式化を行わない場合には，各単位区間の境界ごとにパラメータ v_{fs} および w_s が同一か異なるかを決定することとなる．これは 2^{78} 通りと膨大であり現実的とは言いがたい．

λ は，同一の FD に従う区間の長さを制御するハイパーパラメータである．統計的に最適な λ は，クロスバリデーションや情報量規準を用いて設定することができる．そのように設定した λ を用いれば，最適な区間設定およびそのときの区間別 FD を求めることができる．また，統計的最適性にこだわらなければ，分析者が λ を調整することで実用上扱いやすい区間数にすることもできる．さらに，式 (9) の解は以下の場合を特殊例として含んでいる:

1. λ がきわめて大きいとき，正則化項が卓越するため，車両軌跡が与えられた区間全体に単一の FD を推定する場合に帰着する．
2. $\lambda = 0$ のとき，正則化項は推定に寄与しないため，すべての単位区間に異なる FD を多数個独立に推定する場合に帰着する．

5. モデル選択

定式化ができたところで，いよいよ本稿の主題であるモデル選択について議論する．

(1) 真のモデル

突然だが，そもそも我々はなぜモデル選択を迫られているのであろうか．それは，解こうとしている問題が以下の構造になっているからである．

[6] なお，$\gamma = 0$ は隣接するパラメータの差のみに正則化項を与える方法に帰着する．この考え方の起源は Fusion Regression[30] にある．そこでは時系列上の隣接パラメータを対象としており，Generalized Fused Lasso における $\gamma = 0$ はこれを任意の隣接関係に拡張したものと等価である．

- 我々の観測できない真のモデルからサンプルが得られている．
- そのサンプルだけから真のモデル，あるいはなるべく真のモデルに近いモデルを特定したい．

したがって，仮に我々が真のモデルなるものを完全に把握していれば，観測データの規模や解像度によらずモデル選択という問題は発生しない．手元のサンプルがいかに真のモデルから発生する可能性が低いものであっても，議論の余地がないからである[7]．当然ながら，このような状況はほとんど起こり得ない．

むしろ，我々はほぼすべてのケースで，真のモデルを表現できるとは限らないモデルを想定している．すなわち，真のモデルに対して想定しているモデルの

- 関数形が異なる場合．たとえば，真の FD が指数関数であるのに多項式関数を想定している場合，
- 説明変数が不足している場合．たとえば，真の区間別 FD に対して粗い区間を想定している場合，
- 誤差分布の仮定が異なる場合．たとえば，真の FD では誤差が対数正規分布に従うのに正規分布を想定している場合

などである．

(2) 真のモデルと想定しているモデルの関係

それでも，我々は手元のサンプルを手がかりに真のモデルに近づくことが可能である．この試みこそが，我々が一般的に行っているモデルの推定にほかならない．前節では，関数形，説明変数，誤差分布という個別の項目を列挙した．しかし，モデル選択は，これらの何かが異なる複数のモデル候補から良いものを選ぶ行為であり，具体的に何が異なっているかは重要ではない．そのため，以降ではこれらを区別せずに扱う．

はじめに，真のモデルと想定しているモデルとの理論的な関係を整理する．ここでは，渡辺[31),32)]にならい，以下の 2 つを定義する．なお，本稿の説明は数学的厳密性を欠くため，詳細に興味のある読者は当該参考文献を参照されたい．

1. **実現可能**とは，想定しているモデルのなかに，真のモデルをぴったり表現できるパラメータが存在することである．
2. **正則**とは，想定しているモデルと真のモデルとのカルバックライブラー (KL) 距離を最小化するパラメータが唯一存在し，かつ，そのパラメータにおけるヘッセ行列が正則であることである．

[7] 一例として，偏りのないコインを投げる試行を考えよう．仮に表が 7 回連続で出たとしても，我々はベルヌイ試行というモデルを疑うことはほとんどない．このようなサンプルが得られる確率は 1%未満であるにも関わらず，たまたま珍しいことが起きたと理解するであろう．

したがって，実現可能かつ正則である場合[8]もあれば，実現可能だが正則ではない場合[9]や，正則だが実現可能ではない場合[10]もある．また，実現可能でも正則でもない場合もある．

(3) 正則な場合のモデル選択規準

a) 2 つの情報量規準

さて，まずは正則な場合を考える．我々が推定するモデルに求めたい性質は，おおむね以下だろう．

1. 未知のデータの予測精度が最大のモデル．汎化誤差が最小のモデルという．
2. 真のモデルとの距離が最小のモデル．自由エネルギー[11]が最小のモデルという．

前節の正則の説明は，あたかも真のモデルと想定しているモデルとの KL 距離が確定的に計算できるような書きぶりであった．しかし，本章冒頭で述べたように実際には真のモデルは永遠に未知であり，したがってこれは不可能である．我々にできることは，手元のサンプルだけをもとに，真のモデルと想定しているモデルとの (KL) 距離の分布や期待値を推測することである．同様に，想定しているモデルによる未知のサンプルの予測精度の分布や期待値を推測することである．手元のサンプルは真のモデルからランダムに生成されたものだから，これらは手元のサンプルに依存して確率的に揺れ動く値であり，あくまでも (確率) 分布として推測される．

そしてこの理由により，上記の 2 つのモデルは一般的に異なるものとなる．正則で実現可能な場合，汎化誤差最小のモデルに対応するものが赤池情報量基準 (Akaike Information Criterion; AIC)[33)] 最小のモデル，自由エネルギー最小のモデルに対応するものがベイズ情報量基準 (Bayesian Information Criterion; BIC)[34)] 最小のモデルである．なお，前者が正則だが実現可能ではないには竹内情報量基準 (TIC)[35)] が対応し，後者は実現可能ではなくても BIC がそのまま対応する．具体的に数式で表せば以下の通りで，値が小さいほどそれぞれの規準における良いモデルである．

$$\mathrm{AIC} = -2\ln L + 2\kappa \tag{10}$$

$$\mathrm{BIC} = -2\ln L + \kappa \ln N \tag{11}$$

[8] 真のモデルをぴったり表現できるパラメータが唯一存在する場合．正則で，かつ KL 距離の最小値が 0 といってもよい．
[9] 真のモデルをぴったり表現できるパラメータが複数存在する場合．周期的・点対称的な関数で発生しやすい．
[10] KL 距離を最小化するパラメータが唯一存在するが真のモデルとは異なる場合．
[11] 対数周辺尤度と呼ぶこともある．その場合は符号の正負の定義により最大化または最小化となる．

ただし，κ はパラメータ数，N はサンプル数，L は最大尤度であり，たとえばパラメータの最尤推定量を尤度関数に代入した値を用いれば良い．

b)　クロスバリデーションとの関係

クロスバリデーションとは，手元のサンプルを訓練データとテストデータとに分割し，訓練データで推定したモデルによるテストデータの予測精度を通してモデルの精度評価や選択を行う手法である．特に，1 個のサンプルをテストデータ，残りの $(N-1)$ 個のサンプルを訓練データとし，すべてのサンプルがテストデータを 1 度ずつ経験するように繰り返す方法を leave-one-out cross varidation (LOOCV) という．直感的に明らかなように，サンプル数 N がある程度大きければ，汎化誤差の期待値を求めている作業とほぼ[12]同等である．実際，一定の条件下で AIC は LOOCV の近似となる．このことは，LOOCV の実行には N 回の推定が必要なのに対し，AIC はたった 1 回の最尤推定でこれを近似可能であるという大きなメリットを示している．

c)　過適合に関する補足

一般に，手元のサンプルに過度に適合するモデルは，そうでないモデルよりも汎化誤差が大きくなる (過適合)．したがって，我々はモデルに対して手元のサンプルを必要以上に精度良く再現することを求めないほうが良い．本節の冒頭で挙げたモデルに求めたい性質の 1 つ目はまさにこのことを主張している．そして，過適合しているモデルは当然真のモデルにも近くない．このことは 2 つ目の性質に含まれている．すなわち，AIC も BIC も，過適合を避けるという意味では同じ性質の規準だが，手元のサンプルへの当てはまりと過適合の回避とのバランスの取り方が異なるということである．

d)　統計的検定との関係

帰無仮説と対立仮説を設け，両者の尤度比に基づいた検定を行う手法が一般的に統計的検定として扱われるものである．詳細は避けるが，サンプル数が大きければ，統計的検定によるモデルの選択は，おおよそ AIC と BIC の中間の規準であることが示されている[31)]．

(4)　データの規模との関係

引き続き正則な場合を想定し，サンプル数が増えていく場合にモデル選択がどのように変化するかを考える．

a)　真のモデルを選択する確率

想定しているモデルが真のモデルを実現可能である場合，前節のどの手法を用いても，サンプル数 N が大きくなるにつれて真のモデルを選択する確率が高くなる．ただし，BIC を用いればこの確率が 1 に収束するのに対し，AIC，クロスバリデーションや統計的検定では 1 には収束しないという違いがある．なお，この違いはこれらの規準のモデル選択方法としての優劣を表しているわけではない．

b)　モデルの複雑さ

本章 **(3)a)** で述べた性質を期待する限りは，サンプル数が増えるにつれて複雑な (変数の多い) モデルが選択されやすくなる．その背後にある理由のひとつは，想定しているモデルが真のモデルとは遠いことである．我々は，サンプル数が増えると同一のモデルのパラメータの推定精度が向上することを期待しがちである．しかし，これが成立するのは，想定しているモデルが真のモデルに十分近いときに限られる[13]．そうでない場合には，パラメータの推定精度が向上するだけでなく，より詳細な情報が**みえる**ようになっていく[14]．そのため，サンプルが増えたことにより，それまでよりも複雑なモデルを選択すべきであることが判明することも多い．実用上の議論点は本稿の最後で後述するが，少なくともこの統計的な事実は受け入れる必要がある[15]．

c)　サンプル数は十分に大きいか

サンプル数の増加によって詳細な情報がみえるようになっていくということは，裏を返せば，サンプル数が小さければ本当は存在する差異の情報が埋もれる可能性があるということである．このことを念頭に置くと，これまで行ってきた正則である場合の議論には重要な注意点が残っている．それは，**手元にあるサンプルにおいてモデルが正則であるか**である．

その大前提は，理論的にモデルが正則なことである．実のところ，これが満たされない場合は身近に多く存在する．たとえばある種の混合正規分布は正則ではないことが知られている．そのうえで，ここで問題にしているのは，理論的に正則なモデルから得られたサンプルから計算されるヘッセ行列が正則かどうかである．このとき，サンプル数が十分に大きければ正則になることは言うまでもない．しかし，その十分なサンプル数が，想像よりもはるかに大きな規模であることも珍しくない．サンプル数が不足すると，真のモデルに含まれる説明変数であっても，手元で計算したヘッセ行列のランク落ちの原因になる[16]．ヘッセ行列が正則で

[12] 汎化誤差は N 個のサンプルを用いた次の $(N+1$ 個目の) サンプルの予測誤差 (の期待値) であり，LOOCV は $N-1$ 個のサンプルを用いた次の (N 個目の) サンプルの予測誤差 (の期待値) をサンプルから実際に計算したものである．

[13] シミュレーションを行うと，どんなにサンプル数を増やしても，これらの規準がシミュレーションで設定した真のモデルよりも複雑なモデルを支持することは少ないことが確かめられる．その一端は **6** 章のシミュレーションでも確認できる．

[14] むしろ直感的には自然な発想であろう．それにも関わらず，パラメータ推定精度の向上を期待してしまうのは，パラメータ推定を勉強する際のサンプル数と推定誤差の関係のイメージが強すぎるからではないだろうか．

[15] なお，近年，深層学習の領域において，過剰なパラメータ数のモデルがむしろ過適合を防ぐ現象 (二重降下) が観察されている[36)]．理論的な考察も進展しており，近い将来に従来とは異なるモデル選択指針が生まれる可能性もある．

[16] このとき，最適化計算でヘッセ行列の逆行列が求まらない．

ない場合，正則な場合の規準 (i.e., AIC や BIC) は，求めたいもの (i.e., **(3)a)** で述べた性質) には対応しない．

さらに残念なことに，ヘッセ行列が正則であることは正則な場合の規準を使える必要条件であって，十分条件ではない．元も子もないことを述べると，最尤推定を行ってヘッセ行列の逆行列を求めるだけでは，正則な場合の規準を使えるかどうかは判断できないのである．これを厳密に知りたい場合は，次節のベイズ推定を行い，事後分布が正規分布で近似できることを確認しなければならない．なお念のため，理論的に正則なモデルから十分なサンプル数を得れば，ヘッセ行列が正則になるだけでなく，事後分布も正規分布で近似できる．ただし，そのサンプル数がどの程度大きいかは，試してみないと分からないということである．

また，事後分布が正規分布で近似できないとき，最尤推定量は漸近正規性を満たしていない．本稿では深入りしないが，この条件下では最尤推定量を用いる根拠が薄くなることにも注意しなければならない．

(5)　正則ではない場合のモデル選択規準

それでは，正則ではない場合はどのようにすればよいのだろうか．現状ではベイズ推定を行って事後分布を求めるしかないというのがその回答である[31]．そして，ベイズ推定を行えば，正則ではなくても，また，実現可能ではなくても用いることのできる規準が存在する．ここでは簡潔にその概念を述べるにとどめる．

a)　広く使える情報量規準

AIC の一般化として，広く使える情報量規準 (widely applicable information criterion; WAIC)[37] がある．これは，ベイズ推定の事後分布をもとに計算できる量である．AIC の性質を受け継いでおり，AIC を特殊な場合として含んでいる．したがって，WAIC をモデル選択に用いると，汎化誤差を最小にするモデルを選択する．また，クロスバリデーションの漸近等価であり，サンプル数がどんなに増えても真のモデルを選択する確率は 1 に収束しない．

b)　広く使えるベイズ情報量規準

同様に，BIC の一般化として，広く使えるベイズ情報量規準 (widely applicable Bayesian information criterion; WBIC)[38] がある．これもベイズ推定の事後分布をもとに計算できる量である．BIC の性質を受け継いでおり，BIC を特殊な場合として含んでいる．したがって，WBIC をモデル選択に用いると，自由エネルギーを最小にするモデルを選択する．また，サンプル数が増えれば真のモデルを選択する確率が 1 に収束する．

c)　現状

ベイズ推定を行うべきだとしても，敷居が高い，あるいは計算時間が現実的でないという問題がある．そこでまずは，現にモデルのパラメータをベイズ推定して事後分布を求めているのならば，WAIC や WBIC をもとにモデルを選択することが望ましいという判断になろう．AIC や BIC が最尤推定を一度行えば求まる量であるように，WAIC や WBIC は事後分布を一度作成すれば求まる量だからである．

土木計画学分野での WAIC や WBIC の適用例は，たとえば以下がある: 居住地選択と交通行動のサンプルセレクションモデルの推定[39]，アセットマネジメントにおける劣化モデルの推定[40]，土地利用モデルのパラメータ推定およびモデル選択[41] である．これらはまさしく，もともとベイズ推定を必要とするモデルを扱っている．そのため，自然な帰結として WAIC や WBIC を用いているものと思われる．

(6)　モデル候補が多い場合

これまで，統計的に望ましい規準の議論を行った．その結果，厳密にはモデル候補の個数だけベイズ推定をしなければならないことが分かった．一方で，モデル候補 (たとえば説明変数の候補) が多い場合には，これはまったく現実的ではない．実用に足るか，すなわち実際に計算可能であるかという別の論点が存在する．

前出のスパースモデリングは，まさしくこの問題に対処する強力な方法である．ただし，正則化項のハイパーパラメータの決定方法が十分には確立されていないこと，推定量にバイアスが生じる (不偏ではない) ことが課題として存在する．そこで，本稿の適用 (**7.(1)**) では，Generalized Fused LASSO の結果として得られたモデル候補において選択される変数のみを説明変数として，回帰モデルを再度推定することでこの問題に対処している．そのうえで AIC や BIC が最小となるモデルを検討している．

また，これまでに不偏性の問題を生じない正則化項の投入方法が検討されている．たとえば MCP[42] である．さらに，WAIC を用いて LASSO のハイパーパラメータを最適化する方法も提案されている[43]．

6.　シミュレーション

実データへの適用に先立ち，データ数および手元のサンプルに応じて選択されるモデルが変化することをシミュレーションで確認する．仮想のサンプルを以下のように生成する．

- $s = 1, 2, 3$ の 3 つの単位区間に 2 つの FD を想定し，$s = 2, 3$ は同一の FD に従うものとする．具体的には，$v_{f1} = 90, v_{f2} = v_{f3} = 100, w_1 = -0.75, w_2 = w_3 = -1$ とする．
- 各 FD に従う誤差を加え，擬似的な密度と速度の

表–1 サンプル 10 個のときのシミュレーション結果

区間構成	AIC	BIC	AIC 選択数	BIC 選択数	v_{f1}	v_{f2}	v_{f3}	w_1	w_2	w_3
{1-3}	195.31	198.12	0	0	95.31			-0.87		
{1-2,3}	187.43	193.04	0	0	93.46		101.34	-0.83		-1.02
{1,2-3}	177.05	**182.66**	**3**	**4**	89.41	101.32		-0.72	-1.01	
{1,2,3}	**176.92**	185.33	2	1	89.41	101.31	101.34	-0.72	-1.00	-1.02

選択数以外は 5 回の平均値を示す．太字は AIC・BIC 最小，選択数最大.

表–2 サンプル 100 個のときのシミュレーション結果

区間構成	AIC	BIC	AIC 選択数	BIC 選択数	v_{f1}	v_{f2}	v_{f3}	w_1	w_2	w_3
{1-3}	1938.43	1945.83	0	0	95.10			-0.87		
{1-2,3}	1899.73	1914.54	0	0	93.78		99.33	-0.84		-0.99
{1,2-3}	**1818.93**	**1833.75**	**3**	**5**	90.68	99.88		-0.76	-1.00	
{1,2,3}	1819.03	1841.25	2	0	90.68	100.57	99.33	-0.76	-1.00	-0.99

選択数以外は 5 回の平均値を示す．太字は AIC・BIC 最小，選択数最大.

表–3 サンプル 1000 個のときのシミュレーション結果

区間構成	AIC	BIC	AIC 選択数	BIC 選択数	v_{f1}	v_{f2}	v_{f3}	w_1	w_2	w_3
{1-3}	19412.0	19424.0	0	0	94.99			-0.87		
{1-2,3}	19023.9	19047.9	0	0	93.45		99.63	-0.83		-0.99
{1,2-3}	**18193.2**	**18217.2**	**4**	**5**	90.12	99.81		-0.75	-1.00	
{1,2,3}	18194.2	18230.3	1	0	90.12	100.00	99.63	-0.75	-1.00	-0.99

選択数以外は 5 回の平均値を示す．太字は AIC・BIC 最小，選択数最大.

サンプルを作成する．誤差は標準偏差が 5 の正規分布に従うものとする．

- 値域は $s = 1$ では $10 < k < 90$，$s = 2, 3$ では $10 < k < 110$ とし，$v < 0$ となったサンプルは除外する．

想定するモデルは，区間の分かれ方に応じた 4 通りとする[17]．すなわち，区間構成が s ={1-3}, {1-2,3}, {1,2-3}, {1,2,3}である．そして，各単位区間からサンプルが 10 個ずつ，100 個ずつ，1000 個ずつ得られた場合の 3 通りについて，それぞれ 5 回ずつ最尤推定を行った．

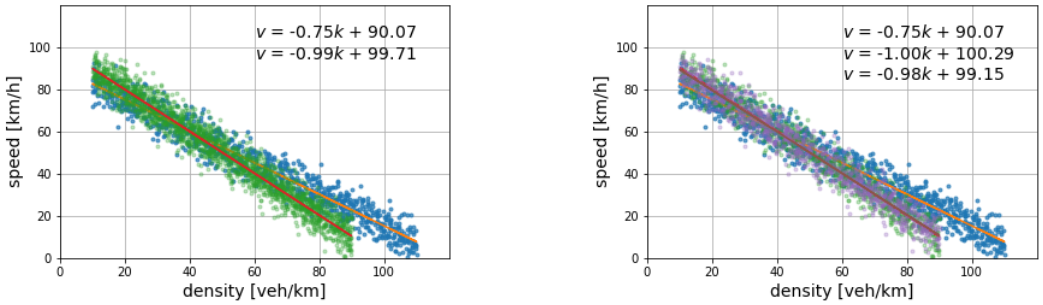
推定結果の平均と，AIC および BIC で各モデルが最適と判断された回数を**表–1-3** に示す．ここから，サンプルが 10 個の時点ですでに真のモデル{1,2-3}を選択する回数が多いことが分かる．ただし，AIC の平均値は真ではないモデル{1,2,3}のほうが小さく，安定した結果ではない．選択結果もサンプルに依存しており，AIC では 2 回，BIC では 1 回真ではないモデル{1,2,3}を選択した．すなわち，今回の設定下では，サンプル 10 個は十分に大きいとはいえないことが分かった．なお，BIC は平均値の意味では真のモデルが最小となっており，ここに AIC と BIC の違いが表れている．

サンプル数を増やすにつれて結果は安定していくが，1000 個まで増やしても AIC は真ではないモデル{1,2,3}を 1 度選択した．これは，AIC は汎化誤差の期待値を小さくできるかわりに，手元のサンプルによってモデル選択が変わりやすいという例である．また，真のモデルがこの例のように比較的単純なときは，サンプル数の増加に応じて不必要に複雑なモデルが選択されることはないということも確認できる．むしろ，誤差のばらつきによりサンプル数が少ないときの方が複雑な (区間数が多い) モデルを選びやすい傾向があった．

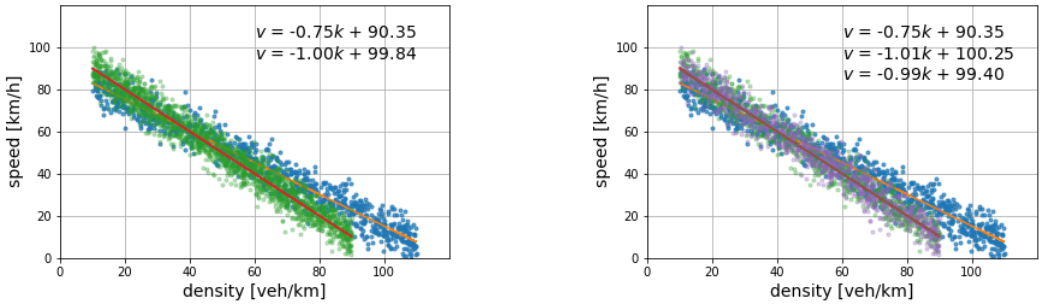
サンプル数 1000 のときの例 2 つを**図–5** に示す[18]．ここで，サンプルおよび回帰直線は区間ごとに色分けしている．また，右の回帰直線のうち単位区間 $s = 2, 3$ の 2 本は重なって描画されている．なお，図の左右でサンプルは同一である．図の上下でサンプルの分布に視覚的な差はほとんどないが，AIC を用いると上では真ではない右の回帰モデルが，下では真である左の回帰モデルが選択される．この例からも，十分大きなサンプル数という量を分析者が想定あるいは判断することの難しさがうかがえる．

[17] ここでは区間が異なれば v_{fs} も w_s も異なるものと仮定した．

[18] なお，**図–3** と**図–5** とを比べるとサンプルのばらつき方が異なっている．現実のデータは密度が中程度の部分の個数が多く，ばらつきも相対的に大きい．厳密にはこの事実を反映した仮想データを作成することが望ましいかもしれない．ただし，一般に行われている「誤算分布に同一の分布を仮定した回帰」は必ずこのような「ずれ」を内包している．データの規模や解像度が高くなる場面では，このことに自覚的になることも必要であろう．



区間{1,2-3}(左) および区間{1,2,3}(右) で回帰した例 (その 1)．AIC では右，BIC では左が支持される．



区間{1,2-3}(左) および区間{1,2,3}(右) で回帰した例 (その 2)．AIC・BIC ともに左が支持される．

図–5 サンプル数 1000 のシミュレーション結果の 2 つの例

7. 実データへの適用

(1) 推定結果

4 章で構築したモデルを **3** 章で作成したサンプルに適用する．ハイパーパラメータ λ を十分に大きな値から徐々に小さくしていくことで，選択される変数が 80 個から 2 個まですべての場合の区間設定を得た．そのうえで，各区間設定において選択された変数を用いて区間別 FD を改めて推定し，AIC および BIC を算出した．**5.(3)** にならうと，本推定での AIC は，区間別 FD から生成される未知 (たとえば将来) の交通状態の予測精度の良さを表す．BIC は，推定された区間別 FD と，(我々が知ることのできない) 真の区間別 FD との距離を表す[19]．

最適な区間数は，AIC では 37，BIC では 21 となった．これらに加えて，区間数が 1 から 3 および 10，40 となる場合を含めた 9 通りについて結果をまとめた (**表–4**)．

また，区間数に応じた区間の分かれ方を**図–6** に示す．これらについて，推定された区間別 FD およびそのパラメータをすべて掲載することは，紙面の都合上困難である．そこで，以下の図示を行う．

- **図–6** に対応する v_{fs} と w_s のヒートマップ (**図–7**)．
- 区間数が 3 以下のときのすべての区間別 FD6 種類 (**図–8**)
- 区間数 21 および 37 のときの推定された FD を重ね合わせたもの (**図–9**)．可読性の観点から回帰直線のみ描画している．

(2) 考察

統計的に最適な区間設定では，40 個の単位区間の多くはそれぞれ単独の区間を構成する (**図–6**)．その場合でも，いくつかの連続する単位区間は同一の区間として推定された．これは，短い単位区間を設定してそれぞれに個別の FD を推定することが，必ずしも地点依存を考慮した最適な FD の推定とは限らないことを意味する．すなわち，本研究のように，同一の FD に従う区間を場所によって異なる長さで推定することに一定の意味があるといえる．たとえば，BIC が最小となる区間数 21 のときは，最長で $s = 24$ から $s = 29$ までの 300 m が同一の区間となる．当該区間は前後をカーブに挟まれた下り勾配の直線区間にあたる．その他，詳細な考察は既報[20] を参照されたい．

(3) サンプル数が小さい場合

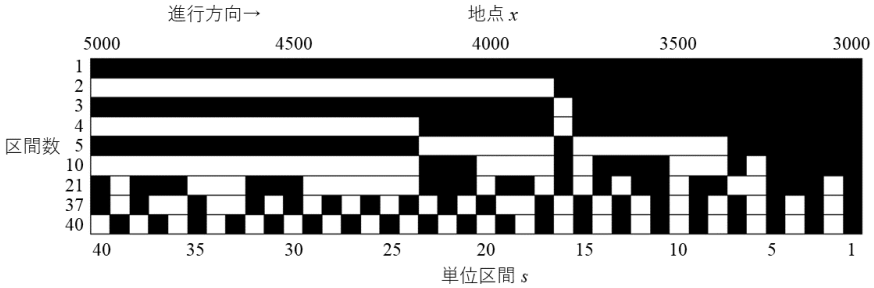
順番が通常と逆ではあるが，データの規模の影響を理解するために，サンプル数を減らして同じ推定を行う．サンプル数を 1/2，1/4，1/8，1/16 とした場合の

[19] 実務上は，現実の区間分割・FD の推定精度を高めたい場合は BIC，現実に発生する交通状態の予測精度を高めたい場合は AIC を用いるのが良い．

表–4 推定結果

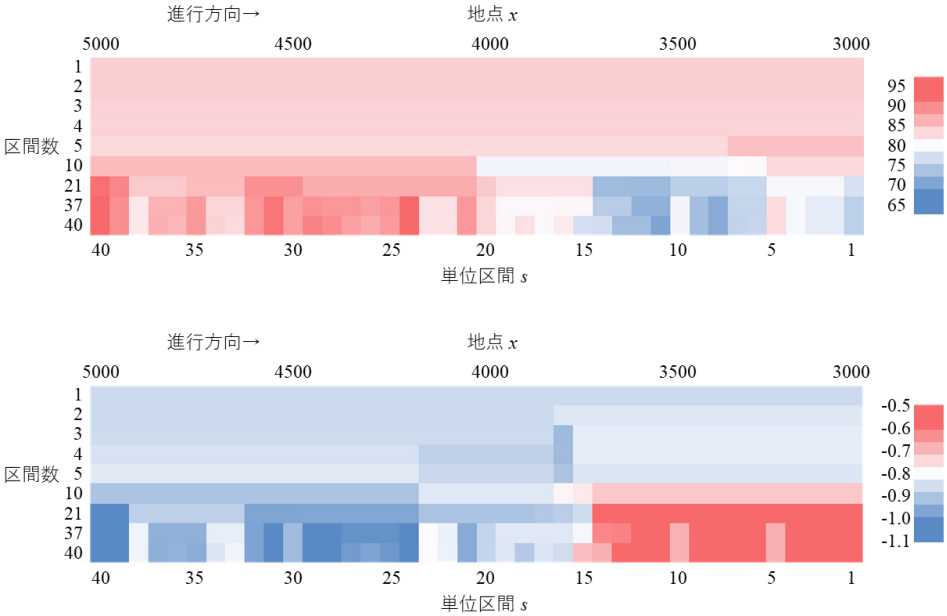
区間数	区間構成 (単位区間 s の組み合わせ)	λ	変数の数	AIC	BIC
1	1-40	18500.292	2	190884.3	190908.7
2	1-16,17-40	16674.702	3	190850.8	190883.3
3	1-15,16,17-40	13765.022	4	190786.2	190826.9
10	1-5,6,7,8-10,11-14,15,16,17-20,21-23,24-40	5889.954	11	189630.7	189728.2
21	1,2,3-5,6-7,8-9,10,11-12,13,14,. . . ,17,18-19,20, 21-23,24-29,30-32,33-35,36-38,39,40	950.489	26	188858.7	189078.3
37	1,2,. . . ,17,18-19,20,21,. . . ,32,33-34,35,36-37,38,39,40	163.36	61	188637.6	189141.8
40	1,2,. . . ,40	0	80	188640.7	189299.5

区間数 21 のとき BIC 最小，37 のとき AIC 最小である．



黒色または白色で連続する部分がひとつの区間を表す．

図–6 区間数ごとの区間構成



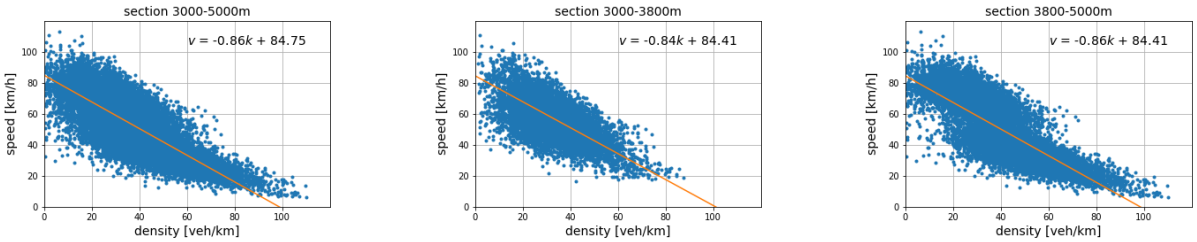
上図: 切片 v_{fs} [km/時]，下図: 傾き w_s [km^2/時]

図–7 区間別 FD のパラメータのヒートマップ

結果を**表–5** に，区間の分かれ方を**図–10** にそれぞれ示す．なお，ここでは，各単位区間ごとにランダムに上記比率となるようサンプルを抽出したものを用いた．

この結果から，情報量規準の性質であるサンプル数が増加すると複雑なモデルを支持しやすい傾向が理解できる．そのうえで，どこまでも複雑になるわけではなく，BIC であればサンプル比率が 1/8 より大きくなるとほぼ安定した結果である．すなわち，仮に真の区間別 FD なるものを想定するならば，Greenshields モデルで表現できる範囲では現状の推定結果はそれを良好に近似できている[20]．一方で，AIC はサンプル比率 1/4 付近で区間数がこの推定で可能な最大値 (i.e., 40)

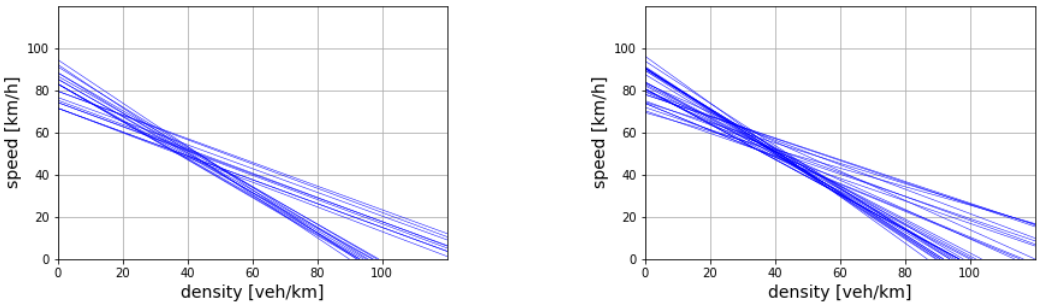
[20] もちろん，本来はベイズ推定を通して正則・実現可能について厳密に議論すべきである．



区間数 1，単位区間 s: 1-40　区間数 2，単位区間 s: 1-16　区間数 2，単位区間 s: 17-40

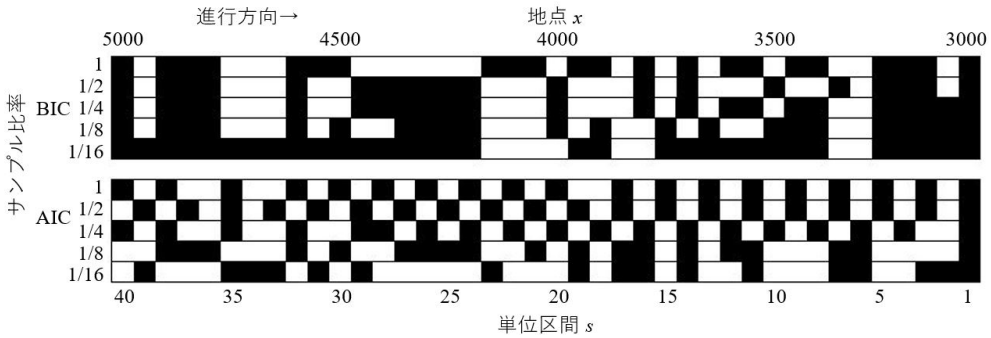
区間数 3，単位区間 s: 1-15　区間数 3，単位区間 s: 16　区間数 3，単位区間 s: 17-40

図–8　区間数 3 以下のときのすべての区間別 FD



区間数 21，BIC 最小　区間数 37，AIC 最小

図–9　区間数 21 および 37 のときの区間別 FD を重ねたもの



黒色または白色で連続する部分がひとつの区間を表す.

図–10　サンプル比率に応じた区間構成

に近づいている．このように，AIC は BIC よりも複雑になりやすく，また汎化誤差と自由エネルギーが異なるものであることも理解できる[21]．

[21] **6** 章のシミュレーション同様に，AIC は汎化誤差の期待値を小さくできるかわりに，手元のサンプルによってモデル選択が変わりやすいということである．区間数がサンプル比率 1/2 で 40 に到達したが，サンプル比率 1 では 37 とやや小さくなるのもこれが理由であると考えられる．

(4)　適用のまとめ

本章では，全車両の軌跡データが得られている道路空間に対して，相互に異なる FD を推定すべき区間を特定しつつ，それらの FD を同時に推定する手法を適用した．1 時間分のデータの推定結果から，対象空間が大きく 2 つの特徴を有する領域に分かれることや，特定

表–5 サンプル数が小さい場合の推定結果

サンプル比率	AIC・区間数	AIC・変数の数	BIC・区間数	BIC・変数の数
1	37	61	21	26
1/2	40	70	21	26
1/4	35	56	19	24
1/8	23	31	22	25
1/16	22	28	7	8

サンプル比率 1 は元データと同一.

の地点にサグあるいは合流の影響が示唆されることが分かった．既述の通り，情報量規準によりモデルや変数を選択すると，一般にデータ数が多ければ多いほど複雑なモデルが選ばれやすい．これが推定された最適な区間数が比較的多かった理由のひとつと考えられる．異なる関数形の導入や定常判定の改良を行えば，より少ない区間数が推定される可能性もある．また，サンプル数を減らした場合の推定結果を含めて検討すると，背後の区間別 FD は区間数 20 程度と考えることができる．一方で，汎化誤差を小さくするという観点では，サンプルが増えるにつれてより小さな区間を支持する傾向があった．

2 章や **5** 章で述べた課題は，そのまま今後の方向性として当てはまる．たとえば，

- 異なる単位区間・領域の大きさや異なるデータを用いること，
- 定常判定手法を改善すること，
- 異なる FD の関数形や説明変数を導入すること，
- FD の関数形も相互に異なる区間別 FD とすること，
- 工学的な精度検証方法を検討すること，
- 正則化項の設定やハイパーパラメータの推定を高度化すること

である．

8. おわりに

本稿では，マクロ交通流の Fundamental Diagram を区間別に推定する方法を例に，データ規模とモデル選択の関係を扱ってきた．**2** 章では FD 推定に大規模で高解像度なデータを用いる際の周辺課題を列挙し，**5** 章ではモデル選択に関わる現状を整理した．また，**6** 章と **7** 章ではデータ規模とモデル選択との関係を実例で示した．まとめにかえて，今後の展開を見据えるにあたって重要だと考える観点を列挙したい．端的には，

- 既知の統計的性質を正しく理解し，
- 現時点で対応可能な点に対応し，
- 現時点で対応不可能な課題の解決を目指す

ということである．

(1) 主義からの脱却

最尤推定やベイズ推定の説明において，頻度主義やベイズ主義という単語が用いられることがある．しかしながら，**5.(4)(5)** で説明したように，それぞれの推定手法の統計的な特性は主義とはまったく関係がない．統計的な主義は，1970 年代に AIC が登場する以前の，合理的に統計モデルを評価するという考え方が存在しなかった時代の遺産である．したがって，現在において主義に基づく議論を行うことは，統計的には意味がないし[22]，実用的には不毛である．

たとえば，ベイズ推定を行う際に事前分布の善し悪しをことさらに心配する傾向があるが，これは情報量規準で確認できることである[23]．一方で，正則ではない場合には，最尤推定よりもベイズ推定のほうが相対的には信頼できる推定結果をもたらす．いずれにせよ，まずは唯一絶対の推定手法など存在しないという認識が大切であろう．

(2) モデルの探索と情報量規準の役割

前節の問題に関連して，情報量規準の名称に惑わされると，最尤推定に対して AIC を，ベイズ推定に対して BIC を用いるというような誤った操作を招きうる．汎化誤差を小さくしたいなら AIC を，真のモデルに最接近したいなら BIC を使うという理解が重要である．もちろんベイズ推定を行っている場合には WAIC や WBIC を使わない理由がない．また，より広い意味で，統計的検定や有意性はあくまでも良いモデルを得るための手段であり，それ自体は目的ではない．近年，交通行動モデルにおける t 値や係数の解釈について警鐘を鳴らす論文[44),45)] が発表された．情報量規準やモデル選択もこの流れに続く必要があろう．

ここからは，実際の問題に対峙する場合の議論である．まず，汎化誤差と自由エネルギーは異なると繰り返し述べてきたが，実際上は自由エネルギーを小さくするモデルを探すことで，ある程度のレベルまでは汎化誤差も小さくなることが多い．その意味では，(W)BIC でモデルを探索することが一つの方法である．

また，正則な場合の規準を，それが適切かどうかは不明ながらも使わざるを得ないならば，少なくともいくつかのモデル候補で AIC や BIC の挙動を確認するべきである．両規準がどちらも良い値を示さない場合には，想定しているモデルに無理がある可能性が高い．その条件下で t 値の意味で有意なモデルが得られたとしても，そのモデルは信頼性を欠くのである．

[22] 講義の導入などで致し方ない場面もあれど，最尤推定はデータが変数でパラメータが定数，ベイズ推定はデータが定数でパラメータが変数という説明にもまったく意味がない．両者は (主義がなければ) 区別がつかず，そもそも区別する必要がない．

[23] もちろん事前分布の設計は重要だが，複数の事前分布の善し悪しは統計的に比較可能であるということ．

(3) サンプル数の影響力

5.(4)c) で説明したように，想定しているモデルが真のモデルに十分接近していない限りは，サンプル数が増えるにつれて複雑なモデルを選ぶことが統計的に支持される．このことは推定時に遭遇する様々な事象をよく表している．たとえば，

- 我々が通常信じる説明変数[24]が有意にならない場合，サンプルが不足している可能性がある，
- 我々が解釈しにくい変数が有意になったとしても，汎化誤差を減らす，あるいは真のモデルに近づくために追加的な変数や別の関数形が必要である，

といったことである．

前者は，仮に想定しているモデルが真のモデルを含んでいたとしても，それでもなおサンプル数が十分に大きくなければ真のモデルが選択されないということそのものである．ただしここで問題となるのは，実社会を相手にする以上，我々が入手できるサンプル数にはそもそも限界があることである．たとえば，仮に全車両の[25]データを集めてもサンプル数が十分でないとすれば，どうするべきであろうか．この問題に対する答えは現時点では明らかでない．後者は次節で検討する．

また，常にサンプル数だけが優先されるわけではない．本稿では扱えなかったが，一例を挙げる．

- 観測精度の悪いサンプルが大量に入手できたとして，それをサンプルとして追加するべきか．
- 観測の解像度を高くすると，モデルに求められる予測の解像度も高くなるため，見かけの精度が落ちる場合がある．

このような場合の対処法は定かではない．類似の問題として，多量の古いデータと少量の新しいデータのどちらを重視すべきか[46] もここに含まれるだろう．

これらは分野特有の問題であり，したがって数理科学分野での発展を待つばかりではなく我々自身が解決していく必要のある研究課題であろう．

(4) 現実的なモデル候補と選択

情報量規準を頼りにモデルを改良していくと，モデルが際限なく複雑になっていくように思われる場合がある．この場合，複雑な順に (W)AIC が支持するモデル，(W)BIC が支持するモデル，現実に解釈可能なモデルの数種類が得られることになる．一方で，我々がモデルに求める性質は，単に将来 (あるいは未観測点) の予測精度だけではない．感度の大きな説明変数を知りたい，実際に制御できる政策変数の影響力を適切に測りたいといった要求がある．これらは統計的な良さとは別の観点で重要である．現実に解釈可能な，あるいは政策に反映可能なモデルでなければ推定する意味がほとんどない場合もある．

そこで，統計的に最適とされるモデルと我々の求めるモデルが異なる場合には，複数の規準や観点でのモデル選択結果を併記することが考えられる．これにより，(W)AIC や (W)BIC でモデル選択を行う場合に比べて偏った推定量や感度になっていないかを，読者も含めたすべての人が確認ができるからである．

なお，本稿で説明してきたモデルの評価方法は，すべてのサンプルを同等に重視することを前提としている．したがって，汎化誤差を議論するにあたって，誤差を小さく予測しなければならないサンプルがある場合や，反対に誤差が大きくても問題が生じないサンプルがある場合は考慮できていない．たとえば，分析対象が社会空間における多様性や公平性を含む場合には，このことは無視できない要素である．こうした場合のモデル探索や選択手法は自明ではない．この点もまた，我々自身に解決が求められている研究課題であろう．

最後に，本稿は決して現状を否定的あるいは悲観的に捉えるものではない．土木計画学の分野では，合理的に考えて一定程度意味のある結果であれば，社会に提示していくことが重要である．だからこそ，我々が扱う問題は真のモデルを正しく想定することがきわめて困難であり，ゆえに統計的な扱いに注意が必要であることもまた事実である．得られたモデルが真のモデルと遠い可能性も，最適でない可能性もあり，しかもどの程度離れているかが不明という状況に対処するのは容易ではない．それでも，一歩ずつ良い方向に進むことが大切であろう．本稿の内容が大規模で高解像度なデータを利用する際の，あるいは新しいモデルを構築する際の参考となれば幸いである．

謝辞: 本稿の内容は，土木学会論文奨励賞の受賞論文「車両軌跡データとスパースモデリングによる区間別 Fundamental Diagram の推定」[20] を出発点としています．同論文に分析・考察を追加し，周辺に存在する課題を整理するとともに，その一部を具体例とともに示したものです．本稿の内容を招待講演として発表する機会をいただき深く感謝申し上げます．受賞論文の執筆にあたって，東京工業大学の朝倉康夫教授 (当時) や瀬尾亨准教授をはじめとする多くの方々にご指導とご助言を賜りました．また，学会や研究会の場では様々な方々に多様な観点からご意見やご指摘をいただきました．分析では，当時の東京工業大学の学生にデータ処理をご協力いただきました．また，本稿で追加したシミュレーションと論文体裁の調整は金子法子氏の尽力により達成されました．ここに記して感謝の意を表します．

[24] たとえば行動モデルやネットワークモデルならば所要時間や料金，FD ならば密度．
[25] 全人口分の，全トリップのなどと置き換えても良い．

利用データである Zen Traffic Data は阪神高速道路よりご提供いただきました．本稿の内容の一部は科学研究費 (若手研究: 18K13846, 基盤研究 (B): 20H02267 および 23H01524) の助成により行われました．

NOTES

注 1) 公益財団法人日本道路交通情報センター: 各種情報の提供 (オープンデータ), https://www.jartic.or.jp/service/opendata/, 2023.10.06 アクセス.

注 2) 一般社団法人社会地盤情報流通推進協議会: インフラデータチャレンジ-道路関係提供データ, https://urbandata-challenge.jp/idc/idc-road, 2023.10.06 アクセス.

注 3) コンペティションの発表資料 (安田昌平, 小倉大輝, 佐々木泰, 福嶋一矢: Gray-Level Co-occurrence Matrix を用いた画像認識による高速道路の渋滞予測, 土木学会インフラデータチャレンジ, 東京, 2019 年 4 月.).

注 4) 阪神高速道路株式会社: Zen Traffic Data, https://zen-traffic-data.net/, 2023.10.06 アクセス.

注 5) 非公開の研究会資料 (朝倉康夫, 吉田光太朗, 中西航, 瀬尾亨: 時空間セルのサイズはプローブカーデータの集計結果にどのような影響を及ぼすか? -阪神高速道路 ZTD データからわかること-, 第 11 回高速道路の交通データ利用に関する勉強会, 北海道建設会館, 北海道札幌市, 2019.09.29.) より.

注 6) 小暮遼河: 全車両軌跡データを用いた交通流基本図のベイズ推定, 東京工業大学修士論文, 2022.

REFERENCES

1) 安田昌平, 池田賢史, 井料隆雅, 石原雅晃: 都市高速道路の需要変動を記述する需要曲線の推定手法, 土木学会論文集 D3, Vol.76, No.5, pp.757-766, 2021. [Yasuda, S., Ikeda, K. Iryo, T. and Ishihara, M.: Demand curve estimation method for describing demand fluctuation on urban expressway, *Journal of Japan Society of Civil Engineers, Ser. D3*, Vol.76, No.5, pp.757-766, 2021.]
2) U.S. Department of Transportation Federal Highway Administration, Next generation simulation (NGSIM) vehicle trajectories and supporting data, ITS DataHub (Data.transportation.gov.), 2016.
3) Seo, T., Tago, Y., Shinkai, N., Nakanishi, M., Tanabe, J., Ushirogochi, D., Kanamori, S., Abe, A., Kodama, T., Yoshimura, S., Ishihara, M. and Nakanishi, W.: Evaluation of large-scale complete vehicle trajectories dataset on two kilometers highway segment for one hour duration: Zen Traffic Data, *2021 International Symposium on Transportation Data and Modelling*, Web conference, 2021.
4) Gloudemans, D., Wang, Y., Ji, J., Zachár, G., Barbour, W., Hall, E., Cebelak, M., Smith, L.and Work, D.B.: I-24 MOTION: An instrument for freeway traffic science., *Transportation Research Part C: Emerging Technologies*, Vol.155, 104311, 2023.
5) 瀬尾亨, 杉本佳昭: プローブカーと感知器のデータに基づくキャリブレーション不要な交通状態推定手法, 土木計画学研究・講演集, Vol.62, Web 会議, 2020. [Seo, T. and Sugimoto, Y.: Calibration-free traffic state estimation using probe vehicle data and detector data, *Proceedings of infrastructure planning*, Vol.62, Web conference, 2020.]
6) Seo, T.: Calibration-free traffic state estimation method using single detector and connected vehicles with Kalman filtering and RTS smoothing, *the IEEE 23rd International Conference on Intelligent Transportation Systems*, Web conference, 2020.
7) 兒玉崇, ペンクレアシュヨアン, 徳永雄斗, 田名部淳: 実測車両軌跡データを活用した相互事故リスクの評価手法の検討, 技術論文, 自動車技術会論文集, Vol.52, No.5, pp.1149-1154, 2021. [Kodama, T., Pencreach, Y., Tokunaga, Y. and Tanabe, J.: Examination of mutual accident risk evaluation method using the vehicle trajectory data observed, *Transactions of Society of Automotive Engineers of Japan*, Vol.52, No.5, pp.1149-1154, 2021.]
8) 吉田光太朗, 中西航, 朝倉康夫: Newell 型モデルのパラメータ推定による時間帯別の車両追従挙動分析, 土木学会論文集 D3, Vol.76, No.5, pp.I_1253-I_1261, 2021. [Yoshida, K. Nakanishi, W. and Asakura, Y.: Analysis of car-following behavior in different time zones by estimating the parameters of Newell's model, *Journal of Japan Society of Civil Engineers, Ser. D3*, Vol.76, No.5, pp.I-1253-I-1261, 2020.]
9) カラジャヤヴェルシニ, 鳥海梓, 大口敬: 高速道路における専用車線からの協調型自動運転車合流機会の評価のためのギャップ分布モデリング, 第 19 回 ITS シンポジウム 2021, Web 会議, 2021. [Kala, J. V., Toriumi, A., and Oguchi, T.: Modelling motorway gap distribution for evaluating merging opportunity for connected-and-automated-vehicles from dedicated lanes, *19th ITS symposium 2021*, Web conference, 2021.]
10) Dahiya, G., Asakura, Y. and Nakanishi, W.: Analysis of the single-regime speed-density fundamental relationships for varying spatiotemporal resolution using Zen Traffic Data, *Asian Transport Studies*, Vol.8, 100066, 2022.
11) Irie, Y. and Akasaka, D.: The study of connected system specification for traffic flow control with effective lane utilization, *International Journal of Automotive Engineering*, Vol.13, No.3, pp.132-138, 2022.
12) 中神勇人, 宇野伸宏: 連続走行軌跡を用いた都市高速道路の速度低下に関する研究, 交通工学論文集, Vol.8, No.2, pp.A_300-A_309, 2022. [Nakagami, H. and Uno, N.: Study on speed reduction on urban expressways by continuous trajectories, *JSTE Journal of Traffic Engineering*, Vol.8, No.2, pp. A_300-A_309, 2022.]
13) 小田紘生, 瀬尾亨, 中西航: コネクティッドカーデータとスパースモデリングに基づく区間別流率密度関数の推定, 第 43 回交通工学研究発表会論文集, pp.513-520, 2023. [Oda, K., Seo, T., Nakanishi, W.: Estimation of location-dependent fundamental diagram based on connected vehicle data and sparse modeling, *Proceedings of the Conference of Japan Society of Traffic Engineers*, pp.513-520, 2023.]
14) 塩見康博: 交通流基本図の変動特性について, 交通工学, Vol.52, No.4, pp.4-10, 2017. [Shiomi, Y.: On fundamental diagram, *Traffic Engineering*, Vol.52, No.4, pp.4-10, 2017.]
15) 和田健太郎, 瀬尾亨, 中西航, 佐津川功季, 柳原正実: Kinematic Wave 理論の近年の展開: 変分理論とネットワーク拡張, 土木学会論文集 D3, Vol.73, No.5, pp.I_1139–I_1158, 2017. [Wada, K., Seo, T., Nakanishi, W., Satsukawa, K. and Yanagihara, M.: Recent advances in kinematic wave theory of traffic flows: variational formulation and network extension, *Journal of Japan Society of Civil Engineers, Ser. D3*, Vol.73, No.5, pp.I_1139–I_1158, 2017.]
16) Haight, F. A.: *Mathematical theories of traffic flow*,

Academic Press, New York, 1963.
17) Daganzo, C.: *Fundamentals of transportation and traffic operations* , Bingley: Emerald, 1997.
18) Seo, T., Bayen, A. M., Kusakabe, T. and Asakura, Y.: Traffic state estimation on highway: a comprehensive survey, *Annual Reviews in Control*, Vol.43, pp.128-151, 2017.
19) Kawasaki, Y., Hara, Y. and Kuwahara, M.: Real-time traffic state estimation on a two-dimensional network by state space model, *Transportation Research Part C: Emerging Technologies*, Vol.113, pp.176-192, 2020.
20) 中西航: 全車両軌跡データとスパースモデリングによる区間別 Fundamental Diagram の推定, 土木学会論文集 D3, Vol.78, No.1, pp.24-33, 2022. [Nakanishi, W.: Sparse modelling for location dependent fundamental diagram with complete vehicle trajectory dataset, *Journal of Japan Society of Civil Engineers, Ser. D3*, Vol.78, No.1, pp.24-33,2022.]
21) Edie, L. C.: Discussion of traffic stream measurements and definitions, J. Almond (Ed.), *Proceedings of the 2nd International Symposium on the Theory of Traffic Flow*, pp.139-154, 1963.
22) Seo, T., Kawasaki, Y., Kusakabe, T. and Asakura, Y.: Fundamental diagram estimation by using trajectories of probe vehicles, *Transportation Research Part B: Methodological*, Vol.122, pp.40-56, 2019.
23) Dahiya, G., Asakura, Y. and Nakanishi, W.: A study of speed-density functional relations for varying spatiotemporal resolution using Zen Traffic Data, *2020 IEEE 23rd International Conference on Intelligent Transportation Systems*, pp.1-8, 2020.
24) Greenshields, B. D.: A study of highway capacity, *Proceedings of the Annual Meeting of the Highway Research Boards*, Vol.14, pp.448-477, 1935.
25) Li, L. and Chen, X. M.: Vehicle headway modeling and its inferences in macroscopic/microscopic traffic flow theory: A survey, *Transportation Research Part C: Emerging Technologies*, Vol.76, pp.170-188, 2017.
26) Tibshirani, R.: Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol.58, No.1, pp.267-288, 1996.
27) Tibshirani, R. J. and Taylor, J.: The solution path of the generalized lasso, *Annals of Statistics*, Vol.39, No.3, pp.1335-1371, 2011.
28) Wang, H. and Rodríguez, A.: Identifying pediatric cancer clusters in Florida using loglinear models and generalized lasso penalties, *Statistics and Public Policy*, Vol.1, No.1, pp.86-96, 2014
29) Inoue, R., Ishiyama, R. and Sugiura, A.: Identification of geographical segmentation of the rental apartment market in the Tokyo Metropolitan Area, *Proceedings of 10th International Conference on Geographic Information Science*, Vol.32, pp.1-6, 2018.
30) Land, S. and Friedman, J.: Variable fusion: a new method of adaptive signal regression, Technical Report, Department of Statistics, Stanford University, Stanford, 1996.
31) 渡辺澄夫: ベイズ統計の理論と方法, コロナ社, 2012. [Watanabe, S.: *Theory and method of Bayes statistics*, Corona Publishing Co.,Ltd., 2012.]
32) Watanabe, S.: *Mathematical theory of Bayesian statistics*, CRC Press, 2018.
33) Akaike, H.: A new look at the statistical model identification, *IEEE transactions on automatic control*, Vol.19, No.6, pp.716-723, 1974.
34) Schwarz, G.: Estimating the dimension of a model, *Annals of Statistics*, Vol.6, No.2, pp.461-464, 1978.
35) 竹内啓: 情報統計量の分布とモデルの適切さの規準, 数理科学, Vol.153, pp.12-18, 1976.
36) 今泉允聡: 深層学習の原理に迫る: 数学の挑戦, 岩波書店, 2021. [Imaizumi, M.: *Approaching the principles of deep learning: a mathematical challenge*, Iwanami Science Library, 2021]
37) Watanabe, S.: Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research*, Vol.11, pp.3571–3594, 2010.
38) Watanabe, S.: A widely applicable Bayesian information criterion, *The Journal of Machine Learning Research*, Vol.14, No.1, pp.867-897, 2013.
39) 渡邉萌, 円山琢也: 多項型の内生的スイッチング構造を持つサンプルセレクションモデル: 居住地自己選択を考慮した交通行動モデリング, 土木計画学研究・講演集, Vol.65, PS1-20, 2022. [Watanabe, H. and Maruyama, T.: A Bayesian sample selection model with multinomial endogenous switching: addressing residential self-selection in travel behavior modeling, *Proceedings of infrastructure planning*, Vol.65, PS1-20, 2022.]
40) 水谷大二郎: 健全度推移の不連続性を考慮したマルコフ推移確率の非集計的推定方法, 土木学会論文集 D3, Vol. 74, No. 2, pp. 125-139, 2018. Mizutani, D.: A disaggregative estimation method of Markov transition probabilities with discontinuous transition of condition states, *Journal of Japan Society of Civil Engineers Ser. D3*, Vol.74, No.2, pp.125-139, 2018.
41) 小林秀佑, 中西航, 堀越光, 高山雄貴: ベイズ推定アプローチによる土地利用モデルのパラメータ推定, 土木学会論文集 D3, Vol.77, No.5, pp.I_469-I_481, 2022. [Kobayashi, S., Nakanishi, W., Horikoshi, H. and Takayama, Y.: Bayesian estimation of a land use model, *Journal of Japan Society of Civil Engineers, Ser. D3*, Vol.77, No.5, pp.I_469-I_481, 2022.]
42) Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, Vol.38, No.2, pp.894-942, 2010.
43) 宮崎大, 渡辺澄夫: 情報量規準 WAIC を用いた LASSO ハイパーパラメータの最適化法, 電子情報通信学会技術研究報告, Vol.114, No.306, pp.213-218, 2014. [Miyazaki, D. and Watanabe, S.: Optimization method of LASSO hyperparameter using WAIC, *IEICE technical report*, Vol.114, No.306, pp.213-218, 2014.]
44) Parady, G., Ory, D. and Walker, J.: The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature, *Journal of Choice Modelling*, Vol.38, 100257, 2021.
45) Parady, G. and Axhausen, K.W.: Size matters: the use and misuse of statistical significance in discrete choice models in the transportation academic literature, *Transportation*, published online, 2023.
46) Sanko, N.: Temporal transferability: trade-off between data newness and the number of observations for forecasting travel demand, *Transportation*, Vol. 44, pp.1403-1420, 2017.

(Received October 13, 2023)
(Accepted ?? ??, ????)

A CONSIDERATION ON SAMPLE SIZE AND MODEL SELECTION: AN EXAMPLE OF LOCATION DEPENDENT FUNDAMENTAL DIAGRAM

Wataru NAKANISHI

The availability of large, high-resolution data is continuously increasing in various areas of infrastructure planning. Model selection criteria used in statistics tend to support more complex models with more variables as the number of samples increases. However, this may not always be consistent with the direction of solution of the problems that our field expects in model estimation. In other words, for our specific problems, it is not clear how the complexity of the model should be captured according to the scale of the data, and what issues exist in such cases. This paper therefore discusses the following, using the estimation of location dependent fundamental diagrams as an example, which has been previously written by the author: the issues related to the data are summarised, the current state of the art of statistical model selection criteria is explained, the properties of the criteria are illustrated with examples, and future perspectives are presented.