植物苗字の分類と地域分布に関する統計分析

塚常 健太・黒川 茂莉（株式会社 KDDI 総合研究所）

日本の苗字はその由来と地域性の点で多様性があり，由来に関する文献学的研究および地域性に関する統計学的研究が行われてきた．しかしながら，苗字の由来を考慮した定量的分析を行っている研究は少ない．本論文では，苗字の由来に関連すると考えられる植物の名前が含まれる苗字（植物苗字）に着目し，その統計学的分析を行う．電話帳に基づく苗字統計の Web サイトより収集した上位1万位の苗字データを用い，漢字辞典を基に植物苗字の分類を行った．その結果得られた1,154種の植物苗字を対象とし，非植物苗字との比較も行いながら地域的な偏りに関する統計的傾向を明らかにした．さらに，その地域的偏りの要因をマルチレベル分析により分析し，植生分布が正の影響を及ぼすことなどが分かった．

Surnames with Chinese characters indicating plants and their regional distributions

Kenta Tsukatsune / Mori Kurokawa (KDDI Research, Inc.)

Japanese surnames have diversity in their origins and regional distributions. Whereas their origins have been studied in philology and their regional distributions have been researched statistically, few studies have focused on the relationship between their origins and regional distributions. In this paper, we focus on surnames with Chinese characters indicating plants (namely, green surnames), and analyzed their statistical characteristics. First, we collected top-10,000 surnames from websites describing statistical information of surnames collected from telephone directories, and annotated green surnames based on a Chinese character dictionary. By analysis of the resulting 1,154 green surnames, we revealed the characteristic skewness of their regional distributions compared with not-green surnames. Then, we applied multi-level analysis to the data of green surnames including their frequency for each prefecture and their variables, and found the positive effect of vegetation related to the plant name in the green surnames on their frequency.

1．まえがき

かつての日本の苗字研究は，歴史上の身分・職階制度や家族制度の変遷を踏まえながら，苗字の発生過程と由来を記述し，分類を行う文献学的な研究が中心であった．しかし，1980年代以降はデータ解析に基づく統計学的な研究も進んできている．例えば，苗字頻度（軒数）と冪乗則との関係性など，統計的性質を解析した研究がおこなわれている[1-6]．また，苗字と地域との関わりに着目した研究としては，日本の苗字マップの作成[7]，苗字ごとの地域的偏在性を数値化した研究[8, 9]，各地域の苗字データを基に人々の移住パターンを推定する研究[10]などが挙げられる．

以上のように，統計的手法を用いた既存研究では，苗字の軒数の統計的性質に加え，その地域ごとの差異に関する知見が蓄積されてきた．しかし，苗字の由来をも考慮した上で定量的分析を行っている研究は少ない．文献学的な苗字研究の成果として，由来の分類では地名由来の苗字が最も多いことが知られているが，苗字に含まれる語彙，例えば植物（の名前）も由来に関係すると考えられている[11]．[11]では「苗字の由来となった植物は，屋敷内に植えられ，家の象徴になっていたとされている．」と指摘されている．

そこで本論文では，統計的分析に反映させることが比較的容易であり，かつ苗字の由来を示す語彙の代表例として，植物の名前に着目する．植物名が含まれる苗字を「植物苗字」と定義し，その地域分布の特徴，さらに分布に対する影響要因について分析を行う．まず軒数上位1万位の苗字のリストを，電話帳に基づく苗字統計の Web サイトより収集し，各苗字について漢字辞典を参照し，植物苗字の分類を行う．次に，各苗字の軒数を関連 Web サイトより収集し，このデータを基に，植物苗字の地域的偏りと影響要因の分析を行う（その都度，非植物苗字の分析結果とも比較する）．

以降，2章では収集した苗字データの内容，3章では基礎統計と地域的偏りの集計結果，4章では植物苗字の地域的偏りに対する統計モデルの分析，5章ではまとめを示す．

2．データの内容

本論文では，以下2種類の電話帳を基に作成された苗字統計に関する Web サイトを利用した．

1) 上位1万位の苗字とその読み仮名，軒数をランキング形式で掲載している「全国の苗字（名字）」[12]　（以降，このサイトから取得したデータを「須崎データ」と略）

2) 任意の苗字の軒数を自治体別に検索できる Web サイト「写録宝夢巣」[13]（以降，このサイトから取得したデータを「写録データ」と略）

まず須崎データを使用し，このランキングに掲載されている範囲の苗字を分析対象とした（下位の同率順位を含むため全部で10,039種）．次に，写録データのうち，前述の上位1万位の各苗字の47都道府県別軒数を使用した．

苗字が植物苗字であるか否かは，植物が由来の漢字・熟語を各苗字が含むかに基づいて判定した．判定の典拠として，漢字の字義解説の中で植物の分類学的記述（「〜科の落葉植物」など）を含んでいる漢字辞典[14]を利用した．辞典において掲載された漢字・熟語の語義が「木」「草」などの単なる植物一般や，「枝」「根」などの部位，「藪」「朽」などの状態，「笠」「蔀」などの加工品ではなく，「松（マツ）」「漆（ウルシ）」など特定の植物種を指している場合のみ，それを含む苗字を植物苗字と定義した．

また別の留意点として，漢字の字義については，漢字自体の成り立ち，および中国における意味が日本で通用している意味内容と異なる場合もある．植物とは異なる地名・道具などの名称を指す場合（例えば「漆」は中国・陝西省の川の名を指す），また植物ではあっても日本とは異なる種類のものを指していたところに，日本独自に別の植物の名称としても定着した場合（例えば「楓」は今日カエデを指す場合が多いが，元々はカツラを指していた）などが挙げられる．しかし，日本の苗字にその漢字が採用される過程では，日本での通例の用法が前提になっていると類推される．そこで，辞典で説明されている字義に植物名（日本独自の場合を含む）があり，それに対応する漢字と読み仮名が共に苗字に含まれている場合は，植物苗字と判定した．

次に，植物苗字の都道府県ごとの分布に影響を及ぼしていると考えられる要因（変数）を検討し，対応するデータの収集を Web 上および書籍から行った．苗字の大半が明治時代以前に出現したという時系列を考慮すれば，本来は遡れる限り過去のデータを使用すべきであるが，入手困難なものも多い．そこで今回はベースラインとして，須崎データおよび写録データと同時期である2010年代以降のデータを使用した（地名や過去の人口分布も含め時系列を考慮した分析は今後の課題である）．採用したデータの内容については，統計モデルの説明と共に後述する．

３．基礎統計

３－１．植物種に関する集計結果

10,039種の苗字のうち，前述の基準で判定を行い，手作業で集計した結果，1,163種（11.58%）が植物苗字であった．また，出現した植物種の数は95種類となった（ただし，須崎データに従い，異体字・旧字なども別種として計上している）．出現した植物名を以下に示す．尾花，李，桑，栗，梨，朴，杠，杉，松，椚，柿，柘植，柏，柚，柳，柳，栂，栃，柾，桜，桧，桂，桃，桐，梅，栢，梛，梶，椎，椋，椙，椛，椿，楠，楳，楓，楡，楢，榎，樺，榛，槙，槇，榊，槻，樅，樗，樫，橘，檜，檀，櫻，漆，瓜，百合，稗，稲，竹，笹，篠，粟，綿，胡桃，芝，芹，芦，茅，荏，茶，荻，菅，菊，菰，菖蒲，菱，葛，萱，萩，蒜，蒲，蓮，蔦，蓬，蓼，蕨，蕪，薄，藤，藍，蘭，蘭，豆，韮，麦，麻，以上である．

なお，植物苗字の候補であったが，最終的に除外したものも存在する．採用基準で植物種か否かの判定が困難な場合，例外的に須崎データの読み方を参考とした．例えば，「薄」という漢字については「薄」姓に「うすい」などの他に「すすき」という読み仮名があるので採用しているが，「薄衣（うすぎぬ）」姓の読みはススキという植物を示しているとは見なしがたいため，こちらの苗字は採用していない．

また，須崎データのランキングに登場する苗字の中で，中国大陸・朝鮮半島の苗字としてもメジャーであるものは，日本の植生分布と対応させて由来を説明することが困難であるため，この後の集計・分析からは除外することとした．漢字の「李」は植物種としてはスモモと見なせるが，「李」姓を除外となるため，後の集計からは除いている（同様の理由で「朴」姓も後の集計からは除外したが，「朴木」姓が別に存在するため，「朴」の字は残っている）．

なお，須崎データに含まれるものの，JIS 外字であったり，異体字を含む一部の苗字は写録データに含まれていないことが判明した．従って，「柳」という漢字も後の分析からは除外となるほか，個別に「髙松」などの苗字も除外となり，最終的に集計・分析に使用するのは植物種93種，苗字種1,154種となった．

表1 苗字ランキング上位の例

	全ての苗字		植物苗字		
順位	苗字	軒数	苗字	軒数	全体順位
1	佐藤	474,558	佐藤	474,558	1
2	鈴木	418,463	伊藤	265,902	6
3	高橋	348,628	加藤	211,491	10
4	田中	327,593	松本	155,284	15
5	渡辺	268,896	斎藤	147,210	17
6	伊藤	265,902	藤田	94,960	30
7	山本	264,577	後藤	93,325	33
8	中村	256,902	近藤	90,170	36
9	小林	250,901	遠藤	81,292	39
10	加藤	211,491	藤井	78,964	41

次に，採用した植物苗字を対象に基礎的統計

を行った．全ての苗字と植物苗字の上位10位を比較すると，植物苗字の上位で「藤」を含むものが9種類を占めていた（表1）．さらに，植物種ごとの苗字の種類数，苗字が占める総軒数（須崎データの軒数），苗字の平均軒数をそれぞれ算出し，上位10位までを棒グラフで示した（図1-3）．これを見ると，すべてのグラフで「藤」が他を引き離す高値を示している一方，種類数では最下位（1種類のみ）であった「柘植」が平均軒数では9位となるなど，同一植物種内部での苗字の軒数に，様々な偏在性が含まれることを示唆する結果となった．なお，僅かながら「藤松」「松竹」など二種類の植物種を含有する苗字も確認された（全部で9種）．

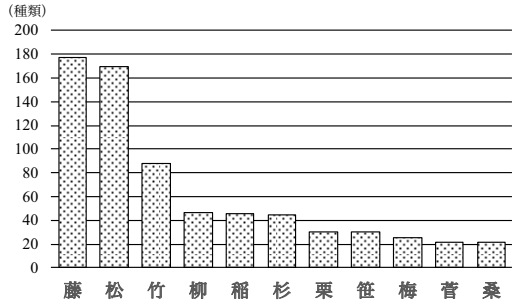

図1 植物別の苗字の種類数ランキング

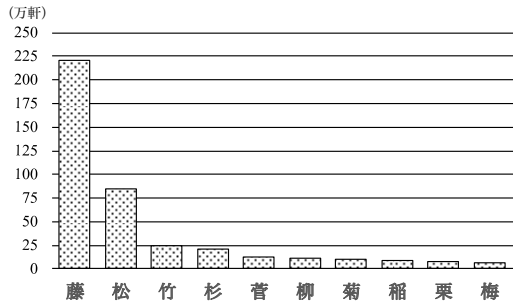

図2 植物別の苗字の軒数ランキング

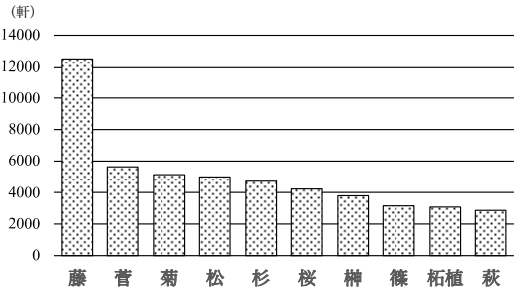

図3 植物別の苗字の平均軒数ランキング

３－２．地域分布に関する集計結果

次に，都道府県ごとの苗字の偏りについての統計的特徴を示す．まず，須崎ランキング上位1万位の苗字のうち，前述の分類基準に従って植物苗字（1,154種）と非植物苗字（8,839種）にデータを区分し，それぞれで都道府県ごとの出現軒数総計を比較した（図4-6）．地図上の分布を比較すると，植物苗字と非植物苗字で類似した傾向が見られるが，東北地方で植物苗字の方が僅かに多いなど，細かい部分で差が存在した．そこで，全ての苗字軒数に占める植物苗字軒数の割合を都道府県ごとに計算し，改めてマッピングした（図7）．その結果，植物苗字が集中する地域とそうでない地域が顕在化した．特に秋田県・山形県など東北地方で植物苗字が集中している一方，西日本では全体的に植物苗字の割合が低く，沖縄県では極めて低水準となっている．例外的に，大分県など，飛び地として植物苗字の割合が高い県も見られる（秋田・山形では各県庁所在地で人口の数パーセントを「佐藤」姓が占めること，九州地方でも例外的に大分で「佐藤」「後藤」姓が多いことなど，「藤」の入る苗字に特に依存している可能性がある）．

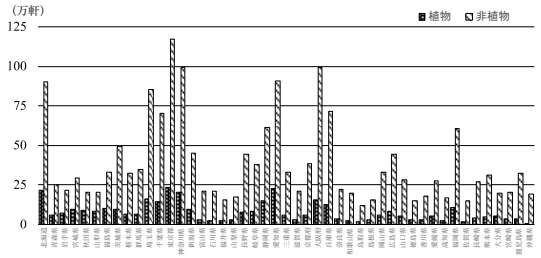

図4 都道府県別の苗字総軒数

３－３．ジニ係数を用いた偏在性の計算結果

次に，苗字ごとの分布の偏在性の高低を示す指標として，ジニ係数（Gini coefficient）の計算を行った．先行研究[8]では，苗字が地域ごとに偏在している状況をジニ係数によって指標化している．そこで本論文でも同様の指標を用いる．まず，苗字それぞれについて47都道府県の分布に関するジニ係数を計算した．具体的には，縦軸に当該苗字の都道府県別軒数の累積割合，横軸に全苗字（写録データ）の都道府県別軒数の累積割合をとり傾きの小さい順に並べたローレンツ曲線と均等配分を示す45度線の間の面積の2倍を計算した．その結果を植物苗字と非植物苗字に分け，散布図で表現したのが図8である（縦軸が写録データに基づく苗字軒数，横軸がジニ係数）．これを見ると，軒数が多い苗字ほどジニ係数が小さく，偏在性が小さくなる傾向が見られ，植物苗字とそれ以外とで大きな傾向は類似しているように見受けられる．また，ジニ係数の平均値を求めると，植物苗字の中での平均値は .627，非植物苗字では .652であり，植物苗字の方が小さく偏在性は大きいことが分かった（Welch の t 検定で $p < .001$で有意差あり）．以上より，軒数の分布に関し，植物苗字，非植物苗字の2群に分けた場合でも，苗字ごとに見た場合でも，差があるという結果を得た．

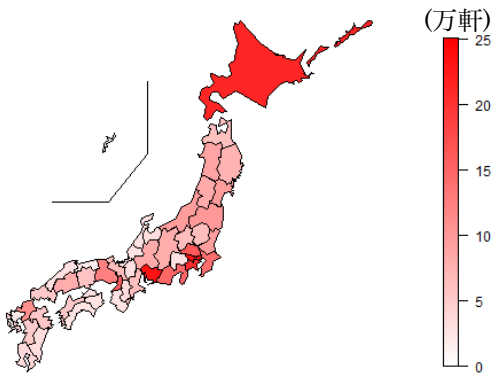

図5 都道府県別の植物苗字の出現軒数

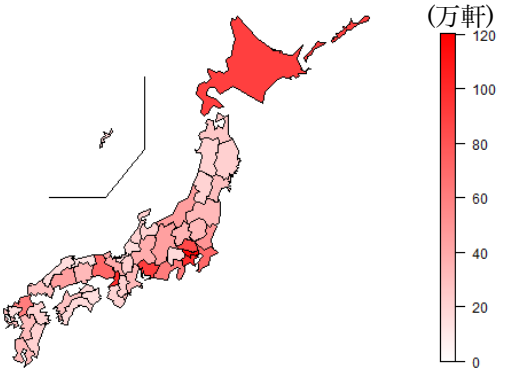

図6 都道府県別の非植物苗字の出現軒数

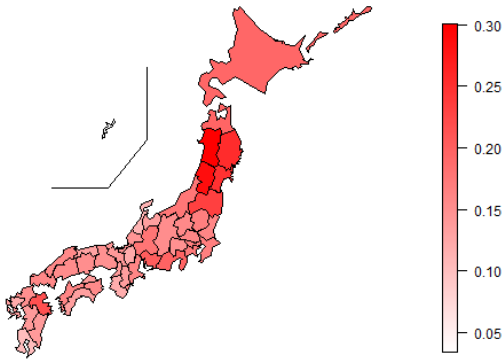

図7 都道府県別の植物苗字の出現割合

最後に，偏在性の高い苗字と低い苗字を，植物・非植物それぞれで例として表2に示す（なお，非植物でジニ係数が最大値（ .985）であった「喜屋」は，須崎データと写録データとで軒数の乖離が大きかったため，外れ値としてこの表からは除外した）．これを見ると，典型的に沖縄発祥，かつ沖縄県でのみ見られることの多い苗字が上位を占めている．

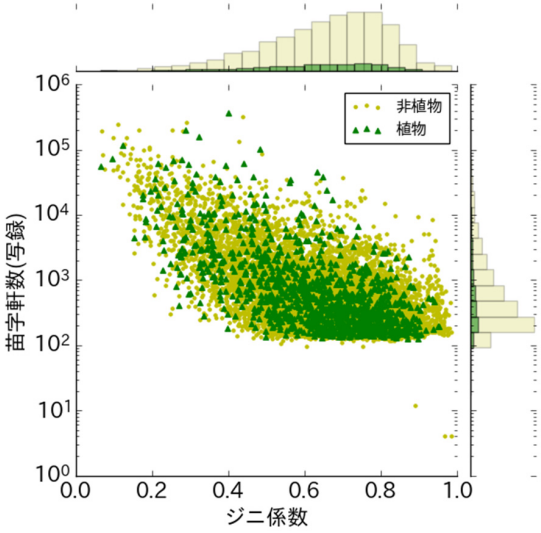

図8 ジニ係数の散布図

表2 偏在性（ジニ係数）上位と下位の苗字

	植物				非植物			
	姓	*Gini*	順位	軒数	姓	*Gini*	順位	軒数
上位	粟国	.951	7931	197	与古田	.984	9507	159
	荻堂	.949	8204	196	仲村渠	.982	4725	445
	稲嶺	.941	4582	463	辺土名	.980	8546	181
	松堂	.938	8370	175	饒平名	.977	6896	293
	稲福	.935	4175	523	根路銘	.976	9278	151
下位	竹田	.175	233	17552	石田	.099	59	49545
	上杉	.152	795	4447	吉田	.098	11	154461
	松本	.122	15	116490	中山	.075	58	50575
	藤田	.094	30	72375	池田	.070	22	84860
	松田	.065	48	55883	中村	.068	8	195219

4．植物苗字の地域的偏りの統計モデル

4－1．説明変数に関するデータ収集

以下では，植物苗字の地域的偏りの要因を分析する．最初に，苗字の分布に影響すると想定される説明変数とそのデータ源，分析に向けた加工のプロセスを合わせて説明する．

まず，直接植物とは関係ないが，コントロール要因として都道府県別の世帯数と固定電話加入率を採用した．絶対的な都道府県別人口の偏りを補正するため，最新の住民基本台帳の都道府県別総世帯数（2018年1月1日時点）を[15] を使用した（説明変数として使用する際には，対数変換を前処理として行った）．また，主に今回使用している苗字データ[12, 13]は電話帳が基となっているが，2018年現在では固定電話加入率が減少傾向にあると考えられ，しかも都市部ほどそれが顕著であると推測される．そこで[16, 17]より最新の都道府県別の固定電話加入件数（それぞれ2018年3月31日時点）を引用し，先述の世帯数で除して固定電話加入率を計算した．

次に，植物に関係する要因として，都道府県

ごとの農業の隆盛度，さらに植物それ自体の繁茂度が影響すると推測されるため，前者については[18]より都道府県別農家数（2018年2月1日時点）を引用して世帯数で除した都道府県別農家割合を計算し，後者については[19]より都道府県別森林率（2012年3月31日時点）を用いた．

また，各地域の植生分布は地理的条件（気候，温度など）にも左右されると考えられるため，それらを代表する変数として[20]より，都道府県庁所在地の緯度・経度・標高を引用した（緯度・経度については十進法に換算）．

さらに，個々の植物の種類によっても日本全国で分布している地域が異なっており，その地域差が植物苗字軒数の地域差とも相関していることが予想される．そこで，植物種に対応する漢字と各都道府県との関係性を表す，植生分布の変数を作成した．まず，植生分布を都道府県単位で説明している大型の図鑑[21-26]（シダ植物のワラビ科のみ[27]）の記述を典拠として，漢字が示している植物種ごとに47都道府県それぞれについて，その県に当該の植物が存在するか否かを「1（ある）」／「0（ない）」の2値データに手作業で置き換えることとした．1つの漢字につき2種類以上，対応する植物種が存在する場合はその和集合を計算し，1種でも植物が存在する都道府県について「1」を割り振ることとした．

なお，典拠となる文献の発行時期について，漢和辞典[14]（1990年代後半）と植物図鑑[21-26]（2010年代後半）との間にタイムラグがあり，この間の植物分類学の進展などの要因から，辞書の記述と異なる系統への転換が行われている種については，[21-26]の記述を優先して判断に用いた（例えば，[14]では「栃（トチ）」=トチノキがトチノキ科となっているが，[21-26]ではムクロジ科トチノキ属トチノキとなっている）．

データの客観性を担保するため，極力例外規則を設けず，以下の基準に応じて漢字に対する植物種の紐づけを行った．

A)漢字が示す名称に対応する単独の種が存在する場合には，その単体の種の分布を採用した．例えば，「蓬（ヨモギ）」はキク科ヨモギ属ヨモギと対応づけた．

B)漢字の示す名称だけでは単体の種を特定できないが，属名（または亜科名）までが特定可能な場合には，その属（または亜科）に含まれている種であり，かつ漢字の示す名称を含む種をデータとして採用した．例えば，「菊（キク）」について，単独の「キク」という植物はないが，キク科「キク」属という階層までは特定できるので，キク属に含まれており，種名にも「キク（ギク）」と含まれる種を全て採用した．例えば，キク属イワギクは採用したが，キク属イワインチンは種名を基準に除外した．なお，「椿（ツバキ）」のツバキ科ではツバキ属の他にヒメツバキ属やナツツバキ属が存在するが，このような類似の名称の属が存在する場合も，漢字と直接対応するツバキ属のみを採用した．

C)属（または亜科）の階層でも特定できない場合は，種名にその植物名が入っているものを採用することとした．例えば，「蘭（ラン）」のラン科では，「ラン属」というそのものを示す属名（亜科名）が存在しない．このような場合，ラン科カキラン属カキランなど，種名に「ラン」と入っているものを全て採用した．なお，種名を基準とするため，属名に「ラン」と入っていないラン科エビネ属ツルランなども採用した．

D)[14][21-27]の双方の記述を照合し，体系的な区別が不可能である場合のみ個別の対応を行った．例えば，[21-27]が野生植物を対象としているため，作物品種として日本で人為的に普及した種類については，部分的に言及されるにとどまっていたり（例えば「蕪」に対応するカブはアブラナ科の概説の中でのみ言及されている），そもそも言及自体がない場合（例えば「柚」に当たるユズはミカン科の記述でも言及がなかった）もある．このよう植物種については，47都道府県全てに植生分布として「1」を割り振った上で，別個に「作物品種ダミー」という変数を作成することで対応した．また，「桐」は[14]の分類学的記述の中ではアオギリ科と説明されているが，実際には二種類の系統の植物を示していると考えられ，キリ科キリ属キリとアオギリ科アオギリ属アオギリのいずれか一方に特定困難であるため，両方を採用した．

以上の分類基準に加え，[14]においてそもそも複数の植物の系統が漢字の字義として記述されている漢字については，須崎データで併記されている苗字ごとの読み仮名を参考にし，最低でも一種類以上の苗字でその植物種に対応する読み仮名が存在する場合にのみ，その植物種（の系統）を採用した．例えば，「椛（カバ，モミジ）」では，カバノキ科カバノキ属の植物群と，ムクロジ科カエデ属の「モミジ」と名のつく植物群，二種類の系統が対応する．須崎データの読み仮名では，「椛」という苗字に「カバ」「モミジ」の両方が存在するため，このような場合にはカバとモミジの両方を対応種とした．このような処理は例外的処理であるため，逆に須崎データでのみ出現するが[14]に存在しないような読みの植物名は採用しなかった（例えば「高椋」という苗字に「タカグス」という読み仮名が存在するが，[14]では「椋」という字に「くす」の読みがないため，「椋」に対してクスノキ科の植物を採用しなかった）．

最後に，ある漢字の異体字や旧字体，また事

実上同じ意味を表す別字（「杉」と「椙」，「梅」と「楳」）については，全く同じ植物種を対応付けた．以上のプロセスを経て，[21-27]から延べ914種の植物種を採用し，そこから和集合を計算して植生分布の変数とした．

上記に加え，「藤松」など二種類の植物名を含む苗字を区別するため「二種類ダミー」という変数を作成した（この変数の必要性については改めて次節で言及する）．

４－２．マルチレベル分析

以下では，都道府県ごとの苗字軒数を目的変数として，それに影響を及ぼす要因を検証する統計分析を行う．

データ構造上，都道府県別の苗字軒数の分散に対しては，複数の水準において影響要因が存在すると考えられる．いずれの苗字種に由来するかという水準（二種類ダミー．作物品種ダミーもここに含むことは可能），いずれの植物種（漢字種）に由来するかという水準（作物品種ダミー），いずれの都道府県に由来するかという水準（ln 世帯数，固定電話加入率，農家率，森林率，県庁緯度・経度・標高），およびいずれの水準にも完全には包含されない水準（植生分布）である．そこで，このような複数の水準からなる説明変数の分散を適切に分析する統計モデルとして，マルチレベル分析（マルチレベルモデル）[28, 29]を採用した．また，目的変数である苗字軒数がカウントデータであり，かつ標本平均よりも標本分散が大きくなる性質を持つことから，通常の線形モデルではなく，負の二項分布を仮定したモデルを採用した．二種類の植物種を含む苗字については，通常のマルチレベル分析では一サンプルが同時に二つ以上のグループに所属することを仮定できないため，機械的に一つめの植物種を割り振った上で，ダミー変数（二類ダミー含むことの効果を測定することとした（例えば「藤松」姓は「藤」グループに所属となる）．

今回はグループ変数（Between）として複数の候補が考えられるため，各サンプル（苗字軒数×47都道府県）の変数のうち都道府県，漢字種（植物種），苗字種のそれぞれをグループ変数として定義した場合の二段階マルチレベル分析を適用し，一段階の回帰分析の結果とも合わせて情報量規準（AIC, BIC）を基に比較し，最も適合度の高い分析結果を採用することとした．グループ変数単位の分散を完全に分析に反映させるためには，都道府県と苗字種を別々のグループ変数とする三段階マルチレベル分析の適用が理想的である．しかし，使用するソフトウェア（Mplus ver.7）では前述の負の二項分布を仮定したマルチレベル分析を二段階までしか実行できないため，今回は三段階の分析を行わなかった．

なお，植生分布，作物ダミー，二種類ダミー以外の変数は都道府県単位で値が決まるため，非植物苗字のデータとも紐づけが可能である．そこで以下では，共通の変数を使用し，対照群として非植物苗字の軒数に対しても同様の分析を行う．都道府県単位の変数群については，植物苗字と非植物苗字で記述統計量および説明変数間の相関係数は同じ値をとる．

表3は分析に使用する変数（グループ変数を除く）の記述統計量，表4はマルチレベル分析に使用する説明変数間の相関係数である．都道府県ごとに決まる変数群（ln 世帯数以下，県庁標高までの7変数）は，植物種ごとに決まる二種類ダミーおよび作物種ダミーとは独立（無相関）となっている．相関係数を見ると，都道府県に関する変数同士では互いに高い相関が見られる．このことに留意した上で，マルチレベル分析結果を解釈する必要がある．

表3 使用変数の記述統計量

	植物 (N=54238)			
	Mean	*S.E.*	*Max*	*Min*
軒数	67.464	524.804	35540.000	.000
ln世帯数	13.642	.807	15.775	12.372
固定電話加入率	.274	.057	.380	.169
農家率	.034	.019	.075	.001
森林率	.628	.151	.840	.305
県庁緯度	35.383	2.591	43.064	26.213
県庁経度	136.040	3.660	141.347	127.681
県庁標高	41.732	72.412	371.300	2.000
植生分布	.966	.181	1.000	.000
二種類	.008	.088	1.000	.000
作物品種	.227	.419	1.000	.000

	非植物※ (N=415433)			
	Mean	*S.E.*	*Max*	*Min*
軒数	43.930	302.638	42353.000	.000

※ln世帯数から県庁標高までの記述統計量は植物苗字と同じ

表4 説明変数間の相関係数

	世帯	電話	農家	森林	緯度
世帯					
電話	-.608 ***				
農家	-.759 ***	.767 ***			
森林	-.644 ***	.594 ***	.530 ***		
緯度	.088 ***	.252 ***	.269 ***	.157 ***	
経度	.282 ***	-.044 ***	.088 ***	-.088 ***	.834 ***
県庁	-.111 ***	.149 ***	.260 ***	.234 ***	.226 ***
植生	-.003	.029 ***	.025 ***	.028 ***	.053 ***
二種	.000	.000	.000	.000	.000
作物	.000	.000	.000	.000	.000

	経度	標高	植生	二種	作物
世帯					
電話					
農家					
森林					
緯度					
経度					
県庁	.334 ***				
植生	.039 ***	.012 **			
二種	.000	.000	.011 *		
作物	.000	.000	.100 ***	-.025 ***	

***: p< .001, **: p< .01, *: p< .05, †: p< .10

以下では，マルチレベル分析本体に入る．二段階の分析を行う際，グループ変数の水準（Between）に対応する説明変数は集団平均，それ以外（Within）の説明変数は全体平均を用いたセンタリングを行った．比較する通常の一段階の回帰分析では，全ての説明変数で全体平均によるセンタリングを実施した．

マルチレベル分析の結果を表5, 6に示す（標準化後の推定値のみ記載）．植物苗字，非植物苗字のどちらにおいても，都道府県をグループ変数とした場合のマルチレベル分析は安定的な推定結果が得られなかったため，表から除外している．また，グループ変数間の傾きに変量効果（ランダム傾き）を含めたモデルもやはり推定が不可能であったため，今回の結果では切片のみに変量効果（ランダム切片）が仮定されている．

表5 植物苗字の分析結果

	1 Lv. (N=54238)		2 Lv. 漢字種 (Between N=93)		2 Lv. 苗字種 (Between N=1154)	
説明変数	*β*	*S. E.*	*β*	*S. E.*	*β*	*S. E.*
ln世帯数	1.172 ***	.061	1.150 ***	.042	1.179 ***	.041
固定電話加入率	.130 *	.059	.127 ***	.036	.176 ***	.041
農家率	.242 **	.088	.148 †	.086	.083	.051
森林率	.135 **	.048	.110 **	.039	.015	.036
県庁緯度	.211 **	.081	.105	.080	.099	.064
県庁経度	.102	.077	.120	.100	-.121 †	.067
県庁標高	-.030	.045	-.002	.033	-.047	.029
Intercepts	5.526 ***	.200				
Means			2.856 ***	.210	.915 ***	.095
AIC	400857.211		392842.043		364945.611	
BIC	400937.322		392931.055		365034.623	

説明変数	*β*	*S. E.*	*β*	*S. E.*	*β*	*S. E.*
log世帯数	.952 ***	.046	1.106 ***	.043	1.160 ***	.038
固定電話加入率	.100 *	.043	.123 ***	.033	.181 ***	.039
農家率	.184 **	.064	.143 †	.083	.059	.048
森林率	.103 **	.035	.104 **	.037	.022	.037
県庁緯度	.150 **	.057	.125 †	.075	.137 *	.068
県庁経度	.077	.056	.092	.098	-.132 †	.073
県庁標高	-.024	.034	.001	.032	-.038	.029
植生分布	.317 ***	.018	.110 †	.065	.105 **	.037
二種類	-.200 ***	.018	-.249 ***	.017	-.051 **	.016
作物品種	-.480 ***	.019	-.054	.100	-.078 **	.026
Intercepts	4.399 ***	.108	2.866 ***	.214	1.700 ***	.091
Residual Variances			.997 ***	.011	.992 ***	.004
AIC	398889.921		392603.413		364169.155	
BIC	398996.735		392719.127		364284.870	

***: p< .001, **: p< .01, *: p< .05, †: p< .10

表6 非植物苗字の分析結果

	1 Lv. (N=415433)		2 Lv. 苗字種 (Between N=8839)	
説明変数	*β*	*S. E.*	*β*	*S. E.*
ln世帯数	1.083 ***	.030	1.083 ***	.016
固定電話加入率	.065 *	.027	.185 ***	.016
農家率	.023	.039	-.040 *	.019
森林率	.062 **	.023	-.007	.014
県庁緯度	.032	.038	-.147 ***	.026
県庁経度	.021	.038	-.031	.027
県庁標高	.020	.020	.004	.013
Intercepts	6.112 ***	.095		
Means			1.869 ***	.015
AIC	2862036.359		2639275.835	
BIC	2862134.793		2639385.206	

***: p< .001, **: p< .01, *: p< .05, †: p< .10

結果を情報量規準から判断すると，植物・非植物共にグループ変数に苗字種を採用した場合の二段階マルチレベル分析が最も適合度が高いことが分かった．また，共に ln 世帯数と固定電話加入率が大きな影響を及ぼしていた．さらに，植物苗字においては二種類ダミーと作物品種ダミーが負の効果を持つ一方，植生分布が正の効果を持っていた．

なお，植物苗字の「苗字種」をグループとした最終モデルについて，説明変数に正規分布とポワソン分布を仮定した場合の分析（線形回帰，ポワソン回帰）も試行したところ（分析結果の表は割愛），それぞれ情報量規準が BIC= 810257.963, 7249595.690となり，やはり負の二項分布を苗字の軒数に仮定する場合（BIC=364284.870）が最適であることが確かめられた．

苗字種をグループ変数とした分析結果に基づくと，農家率，森林率などは植物苗字に影響を及ぼさない一方で，植生分布は正の影響を及ぼしているという結果となった．今回の分析では苗字の由来にかかる（媒介変数など）全ての要因を検証している訳ではないが，植物苗字の分布は少なくとも植生と何らかの関係性があるという示唆が得られた．

5．まとめ

本論文では，植物名が含まれる苗字（植物苗字）に着目し，その地域分布について分析した．Web サイトより収集した上位1万位の苗字データを用い，漢字辞典を基に植物苗字の分類を行った．そのうち1,154種の植物苗字（93種の漢字）を対象とし，非植物苗字との比較も行いながら統計的傾向を明らかにした．

今後の重要な課題としては，まず植物苗字の分布に影響を及ぼす他の要因を考慮することが挙げられる．例えば，植物と苗字とを媒介する有力な中間変数として，（古）地名の影響を考慮することである．また，今回の分析では，地域的偏りに関して言及したように，「藤」という漢字を含むメジャーな苗字に全体の傾向が引きずられている可能性がある．これを「源平藤橘」（藤原氏の影響により，「藤-」「-藤」という姓が多数発生したことが歴史的に知られている）などの特殊なグループとしてコントロールできないか検討することが必要である．さらに，例えば「犬飼」「馬田」といった動物種が入っている苗字グループとの間で異質な特徴が植物種のみで生じるか否かなど，他の語彙的特徴を持つ苗字グループとの比較も有意義であろう．

また，統計的手法に関する課題も残る．今回使用したデータは苗字種（および漢字種）と都道府県に同時にネストされた変数であるが，その二重のネストを適切に調整したモデル（マル

チレベルモデルの拡張版，あるいは全く別系統の分析手法）が可能か検証することも必要であろう．

参考文献

[1] 梅田三千雄. 日本の苗字の計量的分析. 情報処理学会論文誌, 1999, Vol. 40, No. 3, p.796-804.

[2] Miyazima, S., Lee, Y., Nagamine, T. and Miyajima H.: Power-law distribution of family names in Japanese societies. *Physica*, A, 2000, Vol. 278, p.282-288.

[3]佐藤葉子・瀬野裕美. 姓の継承と絶滅の数理生態学 Galton-Watson 分枝過程によるモデル解析, 京都大学学術出版会, 2003.

[4] 千田敏, 間瀬茂. 日本人の名字の統計解析. 日本統計学会誌, 2005, Vol. 35 No. 1, p.55-70 .

[5] 入江治行, 石神英樹, 時田恵一郎. 日本の苗字における多様性と種数面積関係. 日本物理学会講演概要集, 2006, Vol. 61, No. 2-2, p.136.

[6] 早川良, 水口毅. 日本人の名前のサイズ頻度分布. 数理解析研究所講究録, 2012, Vol. 1796, p.26-30.

[7] 矢野桂司. 日本の苗字マップとその応用可能性について. じんもんこん2007論文集, 2007, p.47-54.

[8] 林利充, 大澤義明, 小林隆史. 全国における苗字の空間的偏在とその変化：失われつつある地域性. オペレーションズ・リサーチ : 経営の科学, 2009, Vol. 54, No. 1, p.5-11.

[9] Cheshire, J., A., Longley, P., A., Yano, K. and Nakaya, T.: Japanese surname regions. Regional Science, 2014, Vol. 93, No. 3, p. 539-555.

[10] 齋藤成也. 苗字資料による国内の移住パターン推定の試み. 人類學雜誌, 1983, Vol. 91, No. 3, p.309-322.

[11] 大藤修. 日本人の姓・苗字・名前 人名に刻まれた歴史 (歴史文化ライブラリー 353), 吉川弘文館, 2012.

[12] 須崎春夫. 全国の苗字(名字). http://www2s.biglobe.ne.jp/~suzakihp/index40.html, (参照2018-09-03).

[13] 日本ソフト. 姓名分布&姓名ランキング 録宝夢巣／名前・苗字・名字. https://www2.nipponsoft.co.jp/bldoko/index.asp, (参照2018-09-03).

[14] 小学館辞典編集部(編). 現代漢語例解辞典〈二色刷〉第1版. 小学館, 1997.

[15] 総務省統計局. 住民基本台帳に基づく人口、人口動態及び世帯数. http://www.soumu.go.jp/main_sosiki/jichi_gyousei/daityo/jinkou_jinkoudoutai-setaisuu.html, (参照2018-10-03).

[16] NTT 東日本. 都道県別・事住別加入電話契約数(2017年度末). https://www.ntt-east.co.jp/release/detail/20180531_01_03.html, (参照2018-10-03).

[17] NTT 西日本. 府県別・事住別加入電話契約数(2017年度末). https://www.ntt-west.co.jp/news/1805zpwx/dqch180531a_3.html , (参照2018-10-03).

[18] 農林水産省. 農業構造動態調査. http://www.maff.go.jp/j/tokei/kouhyou/noukou/, (参照2018-10-03).

[19] 林野庁. 都道府県別森林率・人工林率（平成24年3月31日現在）. http://www.rinya.maff.go.jp/j/keikaku/genkyou/h24/1.html, (参照2018-10-03).

[20] 都道府県データランキング. 都道府県庁位置/標高. https://uub.jp/pdr/s/cap_4.html, (参照2018-10-19).

[21] 大橋広好・門田裕一・木原浩・邑田仁・米倉浩司(編). 改訂新版　日本の野生植物1　ソテツ科～カヤツリグサ科. 平凡社, 2015.

[22] 大橋広好・門田裕一・木原浩・邑田仁・米倉浩司(編). 改訂新版　日本の野生植物2　イネ科～イラクサ科. 平凡社, 2016.

[23] 大橋広好・門田裕一・木原浩・邑田仁・米倉浩司(編). 改訂新版　日本の野生植物3　バラ科～センダン科. 平凡社, 2016.

[24] 大橋広好・門田裕一・木原浩・邑田仁・米倉浩司(編). 改訂新版　日本の野生植物4　アオイ科～キョウチクトウ科. 平凡社, 2017.

[25] 大橋広好・門田裕一・木原浩・邑田仁・米倉浩司(編). 改訂新版　日本の野生植物5　ヒルガオ科～スイカズラ科. 平凡社, 2017.

[26] 大橋広好・門田裕一・木原浩・邑田仁・米倉浩司(編). 改訂新版　日本の野生植物　総索引. 平凡社, 2017.

[27] 岩槻邦男(編). 日本の野生植物　シダ　新装版第4刷. 平凡社, 2006.

[28] 小杉考司・清水裕士(編著). M-plus と R による構造方程式モデリング入門. 北大路書房, 2014.

[29] 清水裕士. 個人と集団のマルチレベル分析. ナカニシヤ出版, 2014.