平成 30 年度修士論文

日本人の名前の分布の非一様性に関する解析

大阪府立大学大学院　工学研究科

電子・数物系専攻　数理工学分野

学籍番号 2170103030

鎌田陵平

概要

日本における姓と名の統計的な性質に関する研究を行った。氏名データは市区町村別電話帳二冊と全国都道府県別電話帳を用いた。市区町村別電話帳でそれぞれの自治体の総世帯数・種類数分布とランク・サイズ分布をしらべたところ、総世帯数と種類数の分布においては姓・名ともにべき則の関係がみられた。また、ランク・サイズ分布においてもランクが総世帯数のべき乗でスケールすることで、異なる市区町村のランク・サイズ分布が同じ関数形になった。このようにランク・サイズ分布は区域ごとにべき則が成り立つが、区域ごとの分布の内訳は異なる。この分布の非一様性に着目し、都道府県別の電話帳を用いて区域ごとの名前の分布の差異を調べた。その際着目した統計量は相対サイズであり、その類似性を相関係数で表すヒートマップとデンドログラムで示した。その結果、姓・名ともに地域性があることが判明した。

目次

第 1 章 序論

日本人は姓・名がある。姓は名字、苗字、氏、とも称されるが、本論文では統一して”姓（family name）”と呼ぶことにする。またそれに対して、名は”名（given name）”と呼ぶことにする。姓と名には佐藤や太郎といったありふれているものもあれば珍しいものもあるので、その出現頻度にはばらつきが見られるはずである。そのばらつきにはどういった統計的特徴が見られるのだろうか。

日本人の姓の統計的な性質に関する先行研究としては Miyazima らの愛知県の 5 つの地域の電話帳（1998）のデータを用いて姓の分布について二種類のべき則が成り立つことを示したものや、早川の研究開発支援総合ディレクトリ（Read）を用いて姓・名の分布について調べたものがある[1,2]。Miyazima らは人口と総姓数、姓のランクとサイズの間にべき則が成り立つ事を示している。早川は姓・名ともにサイズと頻度の間にべき則が成り立つ事を示している。外国の姓に関するとものしては Bake,Kiet,Kim らによるものや、Zanette と Manrubia によるものがある[3,4,5]。Bake,Kiet,Kim らはいくつかの国の姓のサイズ頻度分布に関する研究結果を紹介しており、べき則を示す国はべき指数により 2 つのグループに分けられると主張している。

名前以外にも様々な分野でべき則が成り立つ事が報告されている。Dragulescu らによるものや Bettencourt らによるものや様々なべき則についてまとめた Newman によるものがある[6,7,8]。Dragulescu らはイギリスとアメリカそれぞれの国の個人の収入の累積分布が収入が比較的大きい範囲でべき則がなりたつ事を示している。Bettencourt らは都市の人口と都市の様々な指標の間にべき則が成り立つ事を示しており、指標の種類によりべき指数が 3 つのグループに分かれると主張している。

先行研究では、姓の分布に関するものが多く解析対象地域はさほど多くないといえる。しかし、姓と同様に名前も多種多様であり、その分布がどのような特徴を持つかは十分解明されているとは言い難い。また、解析対象地域を増やすと地域による違いという別の側面が出てくることが予想される。以上を踏まえ、本研究の目的は、より多くの地域を対象としたデータを収集し、異なる地域のランクサイズ分布の差異を解析することで、分布の非一様性を特徴付けることである。我々は姓だけでなく名に関しても解析対象として統計的な特徴に関する解析を行った。データは愛知と北海道の市区町村別電話帳 2 冊（2010）と全国都道府県別電話帳（2002）を用いた。

本論文の構成は以下の通りである。2 章では、べき則や姓の統計分布に関する先行研究を紹介する。3 章では、電話帳データを用いて姓・名それぞれの統計分

布についての解析を行う。4 章では出生、死亡、移動のプロセスを考慮し姓の増減についてのモデル方程式を提案する。5 章でまとめと今後の課題について述べる。

第２章　先行研究

本章ではべき則、Zipf 則、Heaps 則についてまとめる。

2.1　統計量の定義

本論文で扱う統計量を表 2.1 にまとめた。

表 2.1 本論文で扱う統計量。

統計量	記号	説明
総世帯数	S	電話帳の総世帯数（$S=\Sigma s$）
総姓数	N	姓の種類数
サイズ	s	同姓の人数
ランク	$r(s)$	サイズで降順に並べた時の順位

総姓数Nからなる総世帯数Sの集団を仮定する。その内訳は表 2.2 のように佐藤 20 人、鈴木 15 人、田中 15 人、鎌田 10 人…であったとする。そのとき、それぞれ姓のサイズsは$s(\text{佐藤})=20, s(\text{鈴木})=15$のようになる。また、ランク$r(s)$は同じサイズの姓があればそれらには同じランクを与える。表の 2.2 では$s=15$の姓が鈴木・田中の 2 つあるので、$r(15)=2$となる。その次のサイズの姓には、重複分を飛ばしたランクを与える。したがって、その次に大きいサイズの姓である鎌田のランクは$r(10)=4$となる。$\varphi_2 \approx 1.33$

表 2.2 総姓数Nからなる総世帯数Sの集団における統計量。

姓	人数	ランク
佐藤	20	1
鈴木	15	2
田中	15	2
鎌田	10	4
⋮	⋮	⋮
希少姓	1	最下位
合計	S	

総姓数N

2.2 様々な分野におけるべき則

この節では様々な分野におけるべき則について紹介する。言語学において、ある本の総単語数と単語の種類数がべき則の関係であることが報告されており、このような経験則は Heaps の法則と呼ばれている。また、単語の出現回数とそのランクもまたべき則の関係であり、Zipf の法則と呼ばれている。これらの法則は他の分野でも報告されている。姓の例において、累積分布$C(s')$とはサイズs'以上の姓の数を示しており、同じサイズsをもつ姓の種類数を$n(s)$とおくと、

$$C(s') = \sum_{s' \leq s} n(s) \tag{2.1}$$

とかける。ランク$n(s)$は同じサイズをsをもつ姓の種類数$n(s)$を用いて、

$$r(s') = \sum_{s' < s} n(s) + 1 \tag{2.2}$$

とも書くこともできる。

以下に 4 つの異なる量の累積分布をしめす。図 2.1～図 2.4 はどの図も両対数でプロットされているのでべき則が成り立っている事がわかる。ただし、図 2.2 は引用数 100 以上においてべき則が成り立つ。表 2.3 にそれぞれの図のべき指数をまとめた。

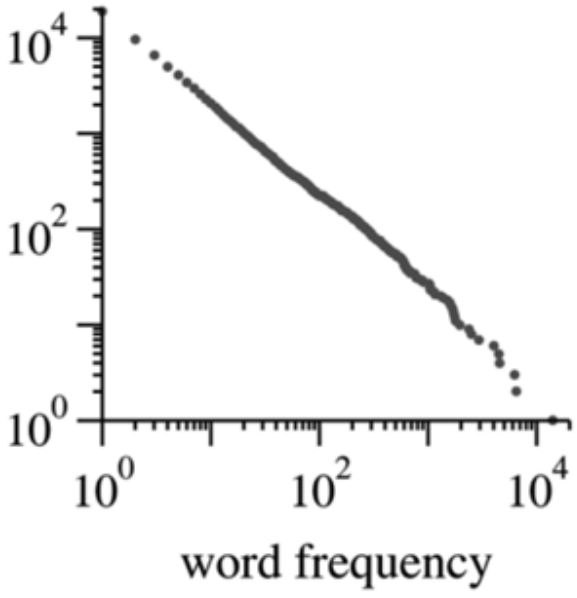


図 2.1 Hermann　Melville 著小説 *Mody Dick* における単語の出現回数の累積分布。（文献[8]）

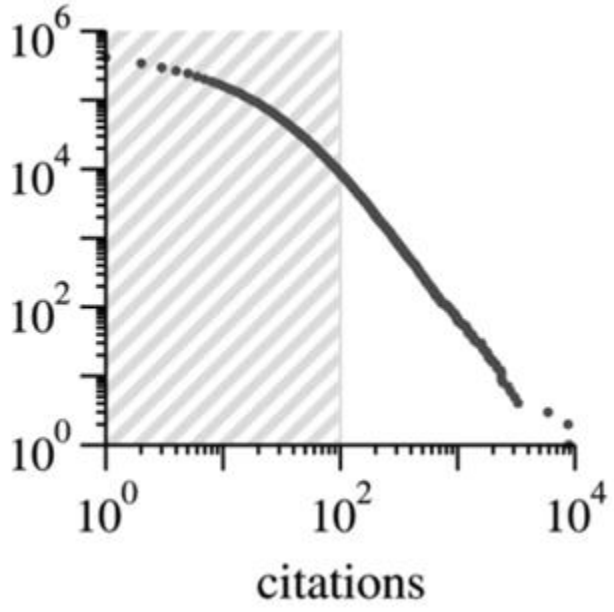


図 2.2 1981 年から 1997 年 6 月までに出版された科学論文の引用数の累積分布。データは Science Citation Index から採集したものである。(文献[8])

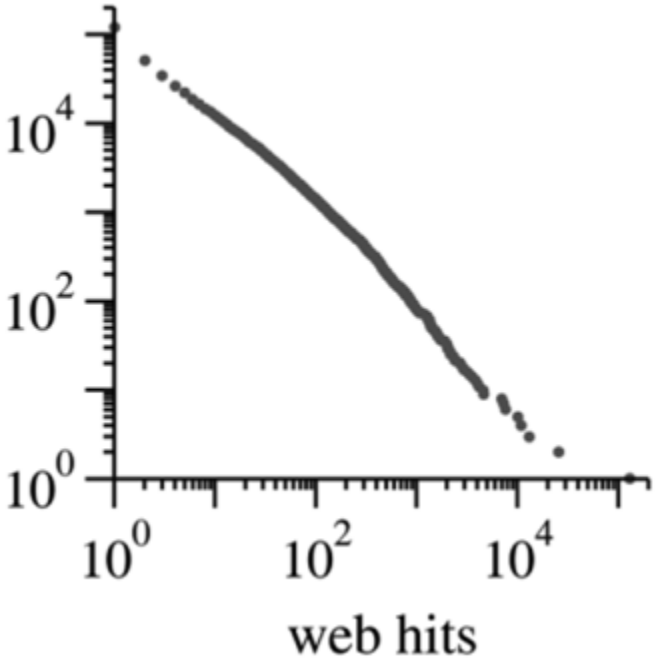


図 2.3 1997 年の 12 月のある一日におけるアメリカのインタネットサービス（AOL Internet service）のユーザー6 万人によるウェブサイトの検索件数の累積分布。(文献[8])

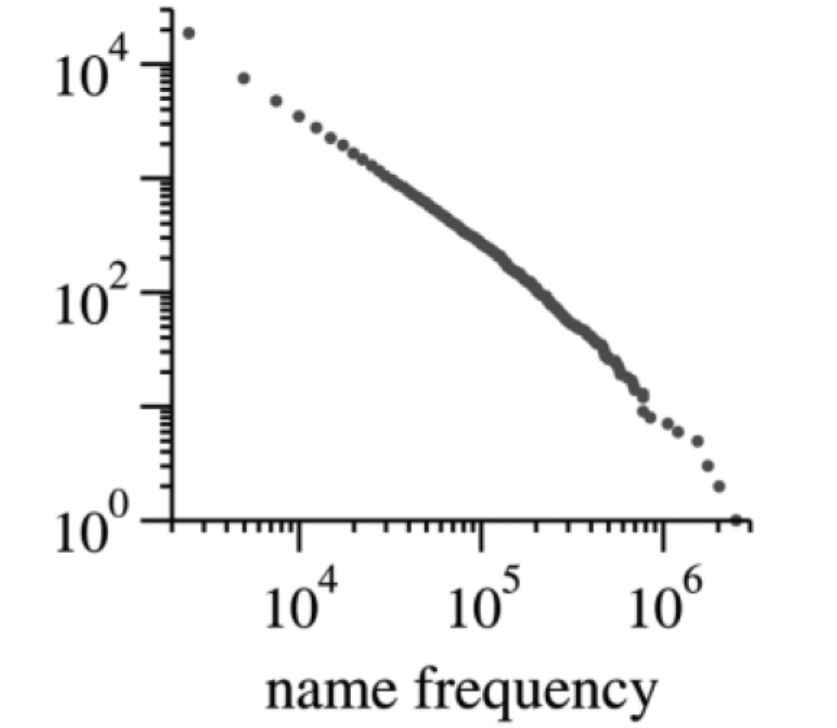


図 2.4 1990 年アメリカにおける最もありふれた姓 89000 個の出現頻度の累積分布。(文献[8])

表 2.3 べき指数のまとめ

	べき指数
図 2.1	2.20
図 2.2	3.04
図 2.3	2.40
図 2.4	1.94

2.3 Miyazima らによる日本人の姓に関する実測結果

Miyazima らは愛知県の 5 つの地域の電話帳を用いて、人口Sと総姓数Nについて解析した。その結果は図 2.5 である。両対数プロットで、傾き 0.65 の直線上に乗っている。これより

$$N \propto S^{\chi} \ , \ \chi = 0.65 \pm 0.03 \qquad (2.3)$$

となり、人口Sと総姓数Nはべき則の関係にある。

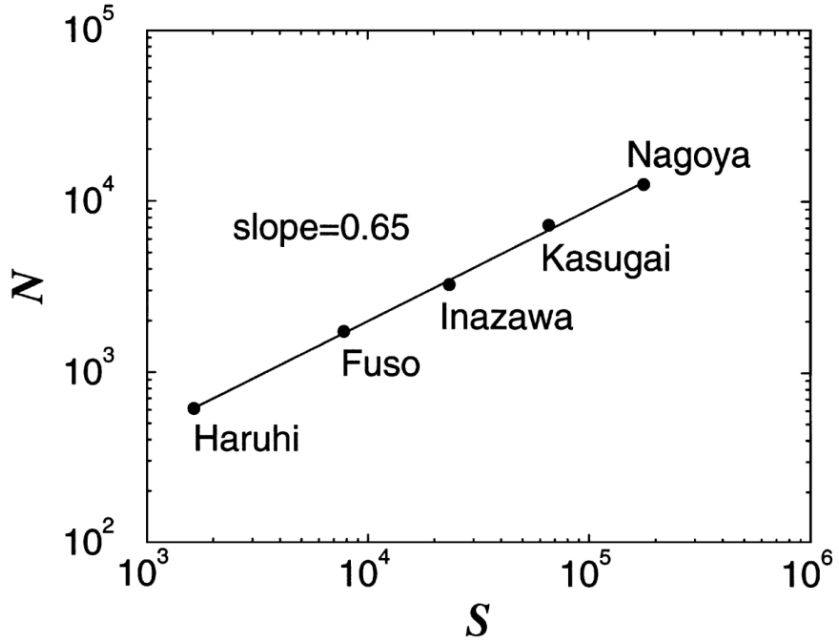
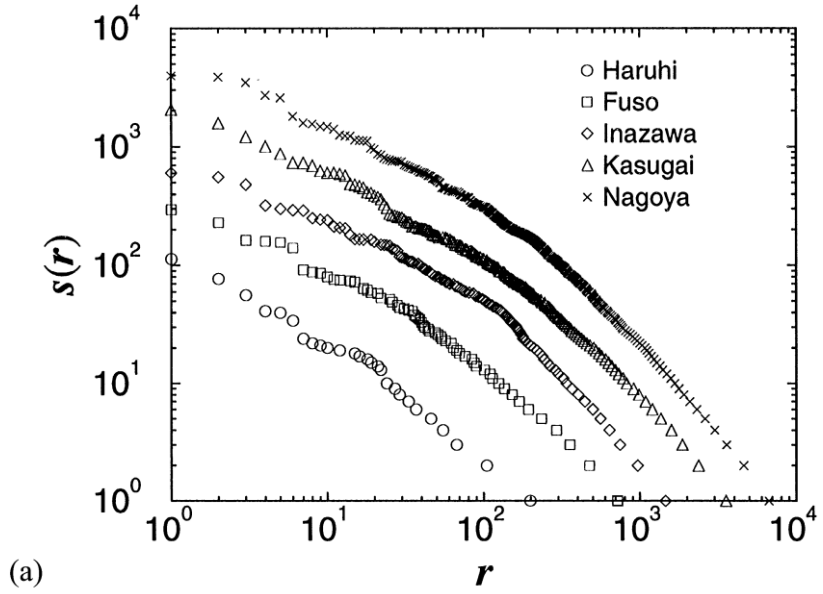


図 2.5 日本の 5 つの地域における総世帯数Sと総姓数Nの関係。両対数プロットして傾き$\chi = 0.65 \pm 0.03$の直線に乗っている。（文献[1]）

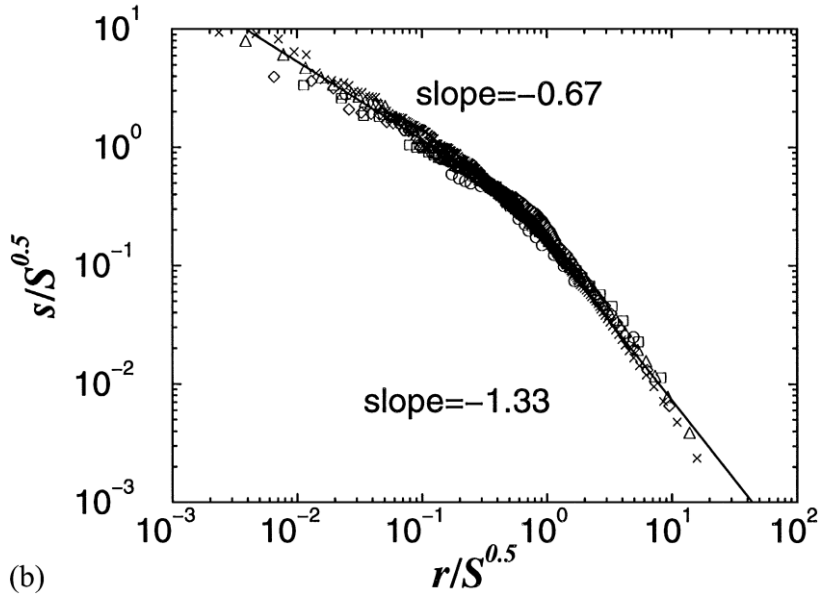
図 2.6(a)はランクrに対してサイズsを両対数プロットしたものである。これをr/S^{α},s/S^{α} $\alpha = 0.5 \pm 0.05$でスケーリングしたものが図 2.6(b)である。スケーリングすることにより 5 つのデータが重なる。また 2 種類のべき則の関係をつなぐクロスオーバーが見られその前後おいて傾きがΦ_1からΦ_2へと変化している。これより彼らは

$$s/S^{\alpha} \propto \begin{cases} (r/S^{\alpha})^{-\Phi_1}, r/r^* \ll 1 \\ (r/S^{\alpha})^{-\Phi_2}, r/r^* \gg 1 \end{cases} \qquad (2.4)$$

を導いている。$\Phi_1 = 0.67 \pm 0.03, \Phi_2 = 1.33 \pm 0.03$である。また$r^*$とはクロスオーバーの起こるランク$r$の値で、$r^* \propto S^{\alpha}$である。式(2.3)は Heaps の法則、式(2.4)は Zipf の法則が成り立つ。



(a)



(b)

図 2.6 (a)両対数プロットでのサイズ$s(r)$のランクr依存性。(b)上のグラフを$r/S^{0.5}$でスケーリングしたもの。ランクの比較的大きい範囲と小さい範囲でクロスオーバーがみられる。その前後の傾きはそれぞれ-0.67,-1.33 である。(文献[1])

2.4 Yule-Simon モデル

べき則を成り立たせるメカニズムを明らかにするためのモデルの一つとして Yule-Simon モデルがある。このモデルは生物学的属の数と属が含む種の数の累積分布がべき則に従う関係（図 2.7）を説明するために、提案されたものである。生物学的種が誕生する進化プロセスの一つに種分化というものがある。ここで扱うパラメータを表にまとめる。

表 2.4 ユール過程で扱うパラメータ

記号	説明
k	種数
m	属が一つ増える間に増える種数
n	属の数
$p_{k,n}$	時刻nで、種数kの属数の割合

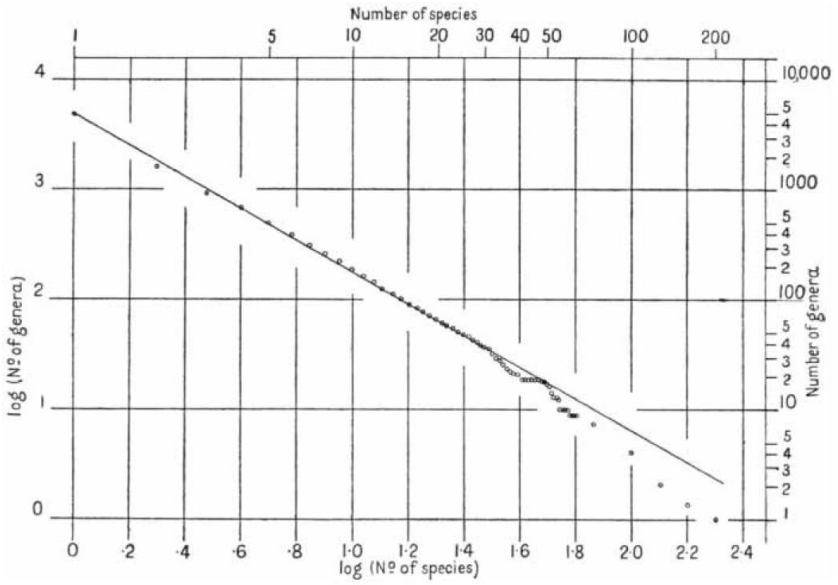


図 2.7 顕花植物の属における種数の累積分布。（文献[8]）

一定の時間間隔で種が分化し、m回ごとに新しい属をつくると仮定する。これを１ステップとして１ステップごとに必ず属の数nは１増える。これより属の数nを時間とみなせる。また、新たに加わるm個の種はそれぞれの属がもっている種の数に応じて分配されるとする。よって、一つの新しい種分化が属i(種数k_i)の中でおこる確率は、

$$\frac{k_i}{\sum_i k_i}=\frac{k_i}{n(m+1)} \tag{2.5}$$

で与えられる。m個の種分化で種数kが一個増える属の期待値は、

$$\frac{k}{n(m+1)} m \cdot np_{k,n} \tag{2.6}$$

となる。上式より時刻nで種数k-1 もつ属が種数kになる、そして種数kもつ属が種数k+1 になる属の期待値はそれぞれ、

$$\frac{m}{(m+1)}(k-1)\cdot p_{k-1,n}$$

$$\frac{m}{(m+1)}k\cdot p_{k,n}$$

となるので、種数kをもつ属数$(n+1)$ $p_{k,n+1}$に対してマスター方程式を書くと、以下の式になる。

$$(n+1)p_{k,n+1} = np_{k,n} + \frac{m}{(m+1)}\left[(k-1)\cdot p_{k-1,n} - k\cdot p_{k,n}\right] \tag{2.7}$$

種数が 1 に対してのマスター方程式は以下の式になる。

$$(n+1)p_{1,n+1} = np_{1,n} + 1 - \frac{m}{(m+1)}\cdot p_{1,n} \tag{2.8}$$

右辺第二項は、毎時刻種数一の属が一つできることを表す。毎回一個の属が一個できるため。)次に、$\mathrm{n}\to\infty$で値$p_k = \lim_{\mathrm{n}\to\infty} p_{k,n}$に漸近すると仮定する。
すると、式(2.8)でp_1について解くと、

$$p_1 = \frac{m+1}{2m+1} \tag{2.9}$$

となる。そして式(2.7)は

$$p_k = \frac{m}{(m+1)}\left[(k-1)\cdot p_{k-1} - k\cdot p_k\right] \tag{2.10}$$

となり、これをp_kについて解くと、

$$p_k = \frac{k-1}{k+1+1/m} p_{k-1} \tag{2.11}$$

となる。式(2.11)をp_1まで展開して式(2.9)を代入すると、以下の式になる。

$$p_k = \left(1+\frac{1}{m}\right)\frac{(k-1)(k-2)\dots 1}{\left(k+1+\frac{1}{m}\right)\left(k+\frac{1}{m}\right)\dots\left(2+\frac{1}{m}\right)} \tag{2.12}$$

ここでガンマ関数$\Gamma(k)$を用いてより単純な式に置き換える。

$$\Gamma(k) = \int_0^\infty t^{k-1}\,exp(-t)\,dt, \Gamma(k-1) = \int_0^\infty t^{k-2}\,exp(-t)\,dt, \tag{2.13}$$

$\Gamma(k)$を部分積分すると、

$$\begin{aligned}\Gamma(k) &= [t^{k-1}\,exp(-t)\cdot(-1)]_0^\infty + (k-1)\int_0^\infty t^{k-2}\,exp(-t)\,dt \\ &= (k-1)\Gamma(k-1)\end{aligned} \tag{2.14}$$

$\Gamma(k)$を(k は自然数であるので) $\Gamma(1)$まで展開すると、

$$\Gamma(k) = (k-1)\Gamma(k-1) = (k-1)(k-2)\Gamma(k-2) = \cdots = (k-1)!\,\Gamma(1)$$

よって式(2.12)の分子は$\Gamma(k)$となる。$(\Gamma(1)=1)$

分母については、以下の式で表す。

$$\frac{\Gamma\left(2+\frac{1}{m}\right)}{\Gamma\left(k+2+\frac{1}{m}\right)} = \frac{\Gamma\left(2+\frac{1}{m}\right)}{\left(k+2+\frac{1}{m}-1\right)\Gamma\left(k+2+\frac{1}{m}-1\right)} = \cdots$$

$$= \frac{\Gamma\left(2+\frac{1}{m}\right)}{\left(k+1+\frac{1}{m}\right)\Gamma\left(k+1+\frac{1}{m}\right)} = \cdots = \frac{\Gamma\left(2+\frac{1}{m}\right)}{\left(k+1+\frac{1}{m}\right)\left(k+\frac{1}{m}\right)\dots(2+\frac{1}{m})\Gamma\left(2+\frac{1}{m}\right)}$$

$$= \frac{1}{\left(k+1+\frac{1}{m}\right)\left(k+\frac{1}{m}\right)\dots(2+\frac{1}{m})} \tag{2.15}$$

またガンマ関数とベータ関数の関係式は以下の式となる。

$$B(k,m) = \frac{\Gamma(k)\Gamma(m)}{\Gamma(k+m)}$$

これより式(2.12)は以下の式に書きかえる事ができる。

$$p_k = \left(1+\frac{1}{m}\right)\frac{\Gamma(k)\Gamma\left(2+\frac{1}{m}\right)}{\Gamma\left(k+2+\frac{1}{m}\right)} = \left(1+\frac{1}{m}\right)\boldsymbol{B}\left(k,2+\frac{1}{m}\right), \tag{2.16}$$

ここでaが大きい範囲において$\mathbf{B}(a,b)$~a^{-b}となるので、

$$\alpha = 2+\frac{1}{m} \tag{2.17}$$

となる。

第３章　電話帳データの解析

先行研究での解析の対象は姓や様々な分野のものであったが、本研究では解析対象を姓と名にした。

3.1 データ

用いたデータは以下の 2 つである。

・2010 年の 愛知県（59 市区町村）と北海道（105 市区町村）の電話帳（以下、それぞれ 2010A,2010H）
・2002 年の都道府県別の電話帳データ（以下、2002All）

それぞれのデータについて姓・名の総世帯数と総種類数を以下の表にまとめる。これは Miyazima らに比べると、対象地域が多く、名のデータもある。早川の ReaD のデータに比べると、2002All は姓・名ともに総世帯数、総種類数が多い。

表 3.1 各電話帳データの姓・名の総世帯数と総種類数

	2010A	2010H	2002All
世帯数	955,412	999,113	29,284,616
総姓数	24,069	23,194	131,481
総名数	74,781	64,694	428,817

姓と名は読み方に関わらず漢字によってのみ区別するものとする。3.2 節では 2010A,2010H の解析、3.3 節では 2002All の解析結果を紹介する。

3.2 北海道と愛知の電話帳データの解析

3.2.1　総世帯数と姓・名の種類数に関する結果

愛知県と北海道の各市区町村ごとの姓の種類数、名前の種類数の総世帯数依存性を調べた。下付添え字f, gはそれぞれ姓と名を表すものである。

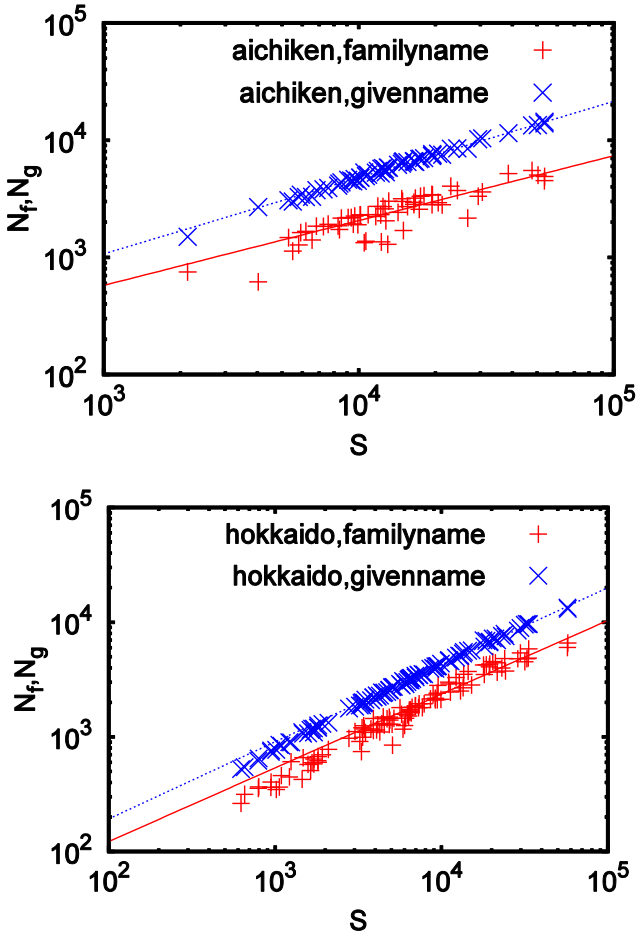


図 3.1.上:愛知県（59 市区町村）における総世帯数Sと姓の種類数N_f(+)、名の種類数N_g(×)の関係。実線と破線は姓、名のデータをフィッティングしたもの。下:同様に北海道（105 市区町村）における関係。

図 3.1 では姓・名いずれの場合においても、両対数グラフで直線に乗っているので、べき則の関係にある。これは Heaps 則が成立していることを示している。フィッティングをする際は最小二乗法を用い両対数グラフで直線近似した。べき指数は直線近似した時の傾きである。式(2.4)のべき指数χにあてはめると、$\chi_f^{Aichi} = 0.55 \pm 0.03, \chi_f^{Hokkaido} = 0.64 \pm 0.01, \chi_g^{Aichi} = 0.65 \pm 0.01,$ $\chi_g^{Hokkaido} = 0.67 \pm 0.01$ となった。

3.2.2　姓についてのランクとサイズに関する結果

愛知県と北海道の各市町村ごとに、それぞれの姓のサイズとランクを調べた。図 3.2 は愛知県の豊田市についてサイズのランク依存性の典型例としてあげる。図 3.2 からrが比較的大きい範囲において、べき則の関係にあることがわかる。

また、図 3.3 は愛知県（59 市区町村）についてのサイズのランク依存性を示している。図 3.3 から愛知県の 59 市区町村においていずれもrが大きい範囲において、べき則関係にあることがわかる。これは Zipf 則が成立していることを示している。サイズを$s_f/S^{\alpha}, \alpha = 0.5$でランクを$r_f/S^{\alpha}, \alpha = 0.5$でスケーリングしたものが図 3.4 である。スケーリングすることにより 59 個のデータが重なり、ランクが比較的大きい範囲において総世帯数によることなくべき則の関係にあることがわかる。また、図 3.5 と図 3.6 は北海道において図 3.3 と図 3.4 と同様にプロットしたものである。北海道の 2 つの図においても同様なことがいえる。図 3.4 と図 3.6 でフィッティングをする際に最小二乗法を用い両対数グラフで直線近似した。べき指数は直線近似した時の傾きである。しかしどこまでがべき則に乗っているかを厳密に定義する事はできない。そのためべき則に乗っている範囲を $2<r_f/S^{0.5}<8$ に限定し、フィッテイングは範囲の最小値を 1 ずつあげていきながら行った。式(2.4)のべき指数Φ_2にあてはめると、$\Phi_{2f}^{Aichi} = 1.3 \pm 0.1, \Phi_{2f}^{Hokkaido} = 1.15 \pm 0.1$となった。

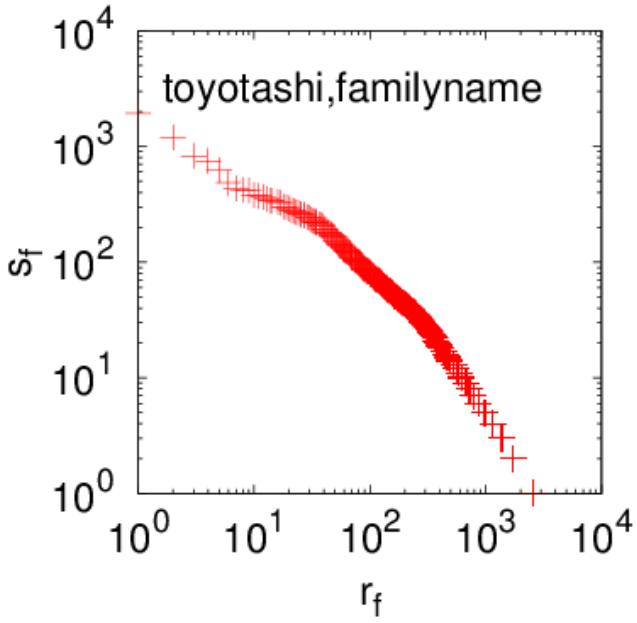


図 3.2 姓についてのサイズs_f（縦軸)のランクr_f(横軸)依存性の典型例（愛知県豊田市、総世帯数 47818 件)。

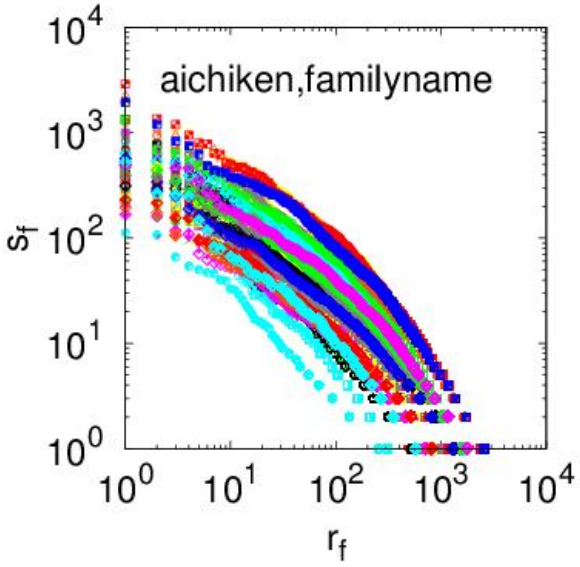


図 3.3 姓についての愛知県の全ての市区町村でのサイズs_f(縦軸)のランクr_f(横軸)依存性。

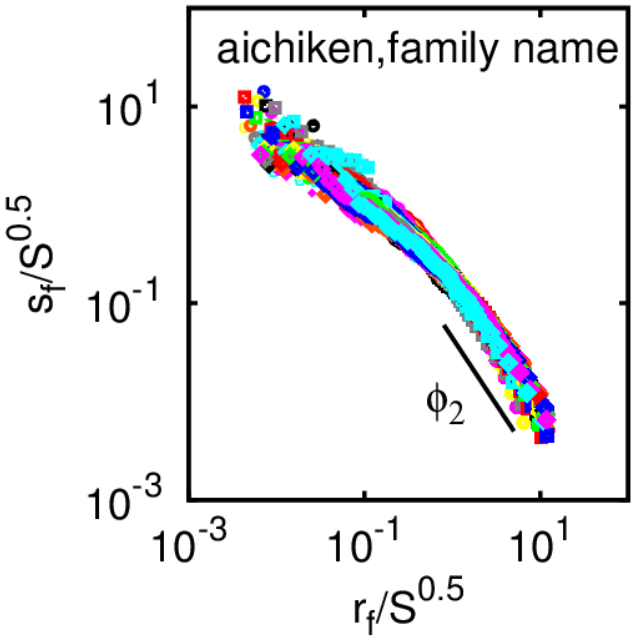


図 3.4 姓についての愛知県（59 市区町村）におけるスケールされたランク$r_f/S^{0.5}$(横軸)とサイズ$s_f/S^{0.5}$(縦軸)の関係。

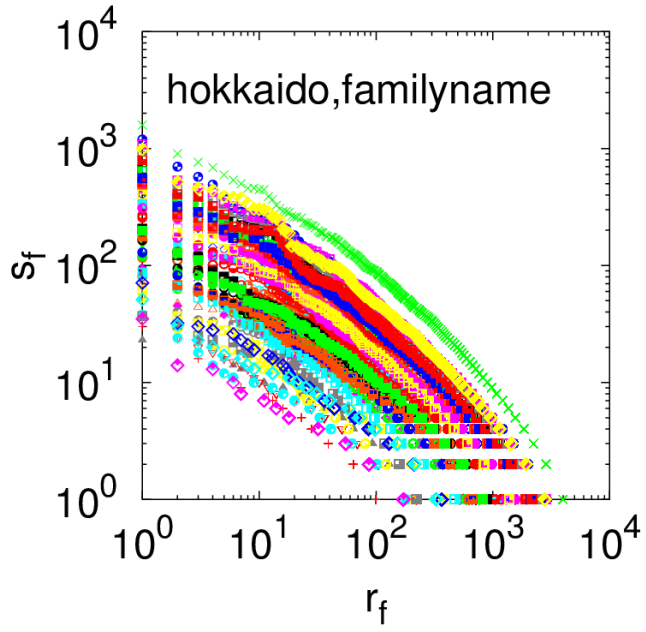


図 3.5 姓についての北海道の全ての市区町村でのサイズs_f(縦軸)のランクr_f(横軸)依存性。

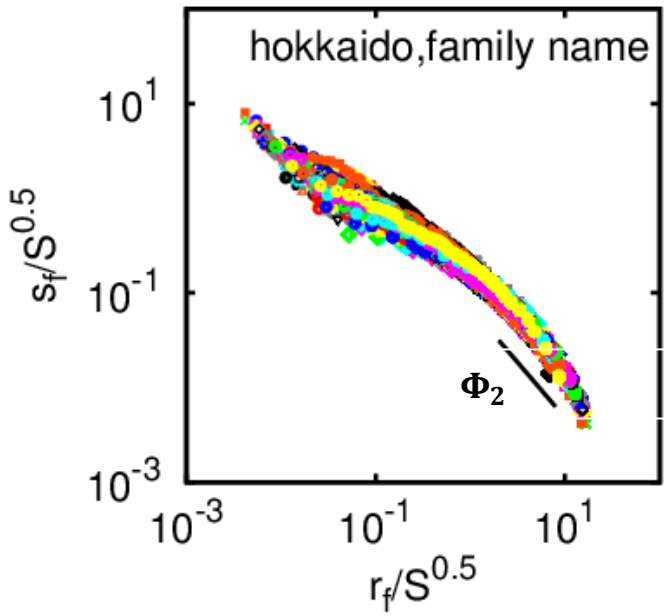


図 3.6 姓についての北海道（105 市区町村）におけるスケールされたランク$r_f/S^{0.5}$(横軸)とサイズ$s_f/S^{0.5}$(縦軸)の関係。

3.2.3 名についてのランクとサイズに関する結果

前小節 3.2.2 と同様に愛知県と北海道の各市町村ごとに、それぞれの名のサイズとランクを調べた。ただし、電話帳データを用いている事から主として世帯主の名前が登録されており、比較的年配の男性の名前で登録されている件数が多いといったバイアスがかかっている。スケーリングについてはサイズを$s_f/S^{\alpha}, \alpha = 0.7$でランクを$r_f/S^{\alpha}, \alpha = 0.5$でスケーリングした。名についても姓と同じくスケーリングすることにより北海道,愛知県のデータが重なり、ランクが比較的大きい範囲において総世帯数によることなくべき則の関係にあることがわかる。図 3.8 と図 3.10 は図 3.4 と図 3.6 と同様な方法でフィッテイングを行った。式(2.4)のべき指数Φ_2にあてはめると、$\Phi_{2g}^{Aichi} = 1.0 \pm 0.1, \Phi_{2g}^{Hokkaido} = 1.0 \pm 0.1$となった。

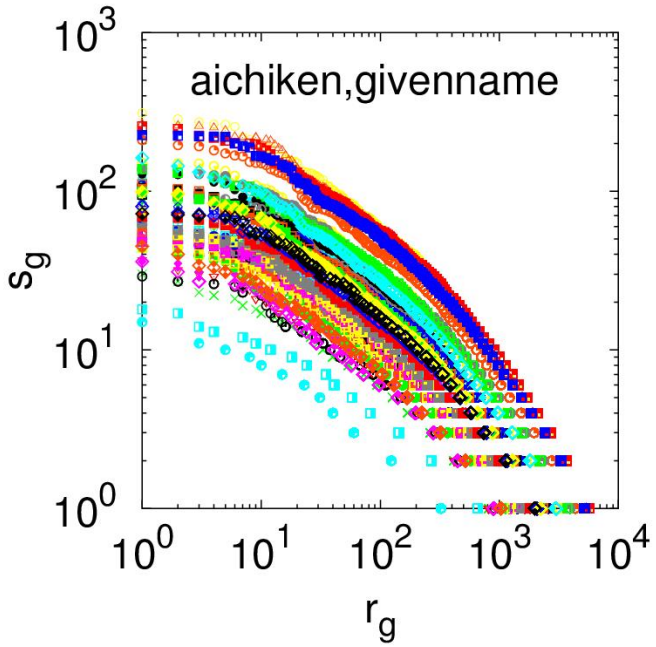


図 3.7 名についての愛知県の全ての市区町村でのサイズs_g(縦軸)のランクr_g(横軸)依存性。

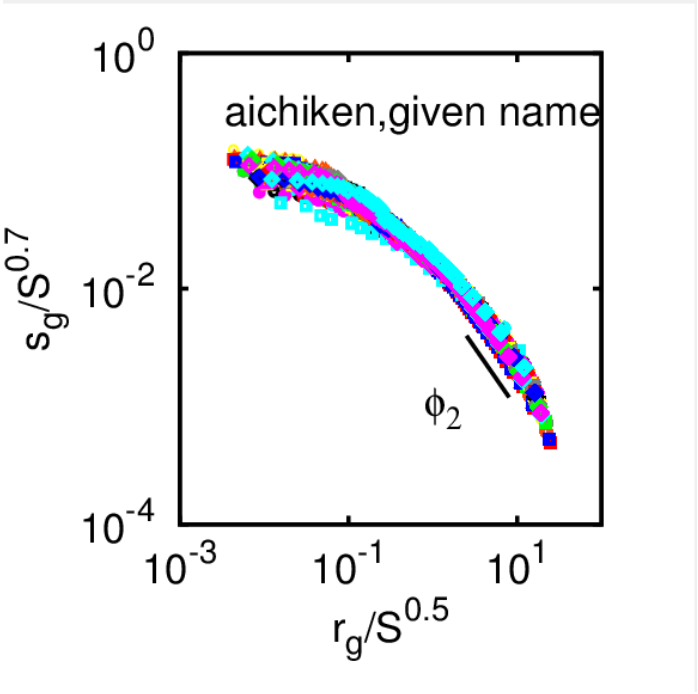


図 3.8 名についての愛知県（59 市区町村）におけるスケールされたランク $r_g/S^{0.5}$(横軸)とサイズ$s_g/S^{0.7}$(縦軸)の関係。

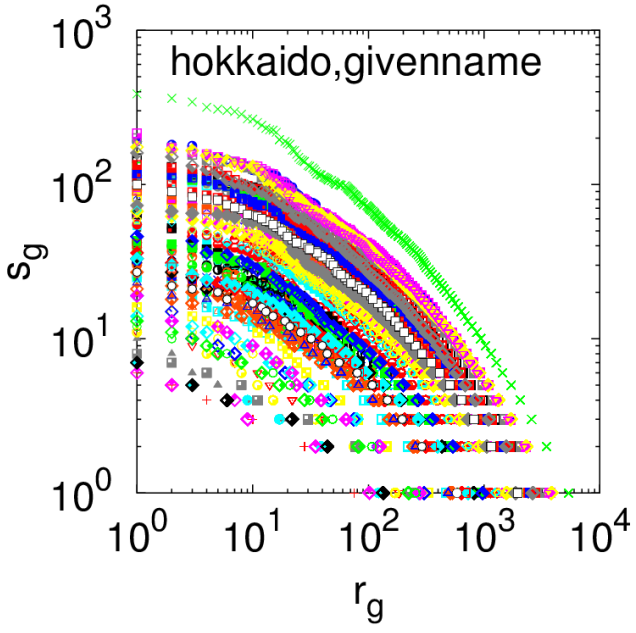


図 3.9 名についての北海道の全ての市区町村でのサイズs_g(縦軸)のランクr_g(横軸)依存性。

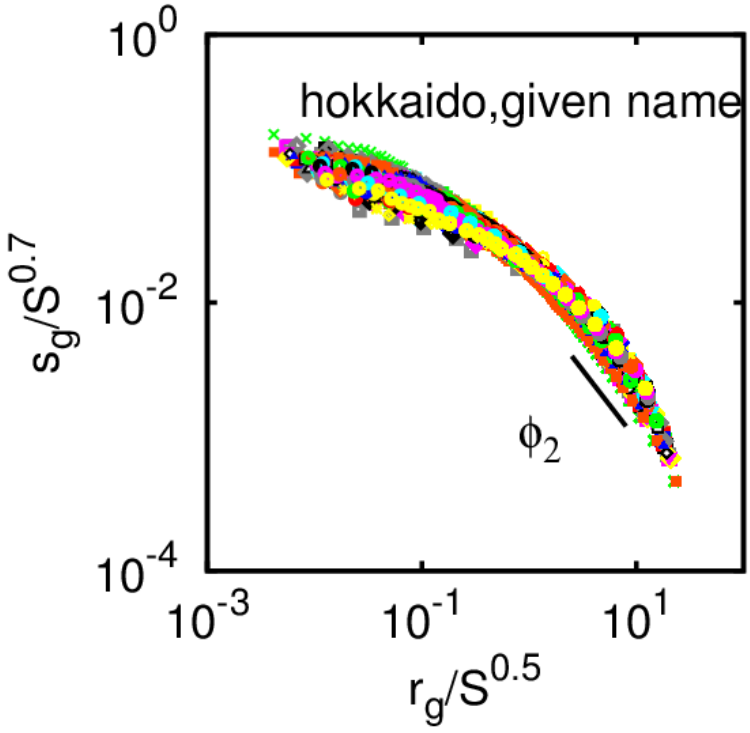


図 3.10 名についての北海道（105 市区町村）におけるスケールされたランク $r_g/S^{0.5}$(横軸)とサイズ$s_g/S^{0.7}$(縦軸)の関係。

表 3.2 べき指数のまとめ

	地域	χ:Heaps 指数	Φ_2:Zipf 指数
先行研究	愛知県内の 5 市区町村	0.65±0.03	1.33±0.03
本研究(姓)	愛知県内の 59 市区町村	0.55±0.03	1.3±0.1
	北海道内の 105 市区町村	0.64±0.01	1.15±0.1
本研究(名)	愛知県内の 59 市区町村	0.65±0.01	1.0±0.1
	北海道内の 105 市区町村	0.67±0.01	1.0±0.1

表 3.2 は各地域の姓と名のグラフにおける指数をまとめたものである。これから Heaps 指数χで姓については若干$\chi_f^{Aichi} < \chi_f^{Hokkaido}$であった。また名については$\chi_g^{Aichi} \approx \chi_g^{Hokkaido}$であった。Zipf 指数$\Phi_2$で姓については若干$\Phi_{2f}^{Aichi} > \Phi_{2f}^{Hokkaido}$であった。また名については$\Phi_{2g}^{Aichi} \approx \Phi_{2g}^{Hokkaido}$であった。

3.3 全国の電話帳データの解析

3.3.1 姓についての分布の解析結果

3.2 節では愛知県と北海道で姓と名ともにランクサイズ間で同様なべき則がみられた。図 3.11 では 2002All を用いて愛知と大阪の姓のランクサイズ分布を示している。指数φの微妙な差異を除けば、両者はランクの大きい領域で同様のべき則に従うと言えよう。しかし、「鈴木」のランクは愛知では 1 位であるのに対し、大阪では 26 位であることからもわかる通り、分布の内訳は異なっている。本章ではこの分布の非一様性に着目する。日本全国を 47 都道府県に分割し、区域ごとの名前の分布の差異を解析することで、その地域における分布の非一様性を特徴付ける。表 3.3 では 2002ALL において全国と最も世帯数が多い東京と最も世帯数が少ない鳥取の総世帯数と総姓数と総名数をまとめた。また、付録 A に全国と各都道府県の総世帯数と総姓数と総名数をまとめた。今回着目した統計量は相対サイズである。サイズではない理由はサイズだと都道府県ごとで量が大きく異なるが相対サイズであれば総世帯数で割り規格化することで比較が容易になるからである。ランクではない理由はランクだと他方の県にしか存在しない名前のランクをつけるのが難しいからである。

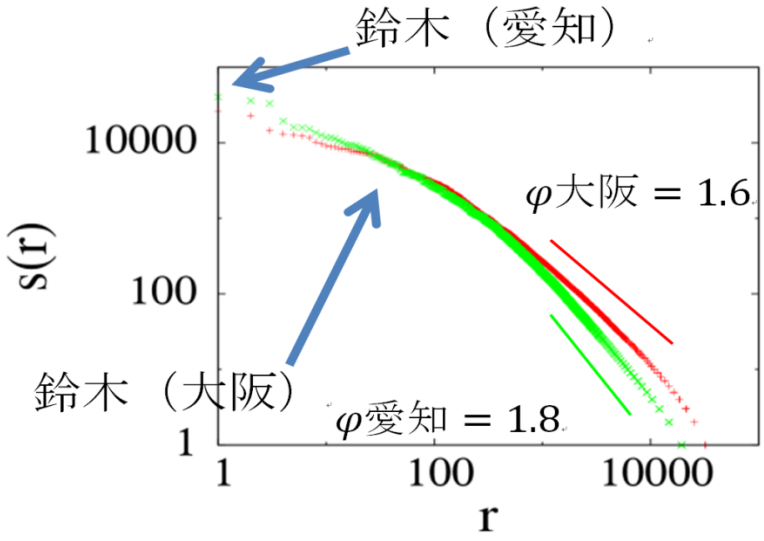


図 3.11 愛知（緑）と大阪（赤）における姓のランクrとサイズsの関係。

表 3.3 全国、東京、鳥取における世帯数、総姓数、総名数

	全国	東京	・・・	鳥取
世帯数	29,284,616	2,103,225	・・・	165,393
総姓数	131,481	45,868	・・・	8,678
総名数	428,817	109,154	・・・	27,165

図 3.12 では大阪と愛知の各姓の相対サイズ（赤点）、大阪と沖縄の各姓の相対サイズ（青点）の関係を示しており、赤点（愛知）は青点（沖縄）に比べて直線y = x付近にあることがわかる。これより大阪と愛知の組み合わせのほうが大阪と沖縄の組み合わせよりも分布の内訳が似ているといえる。

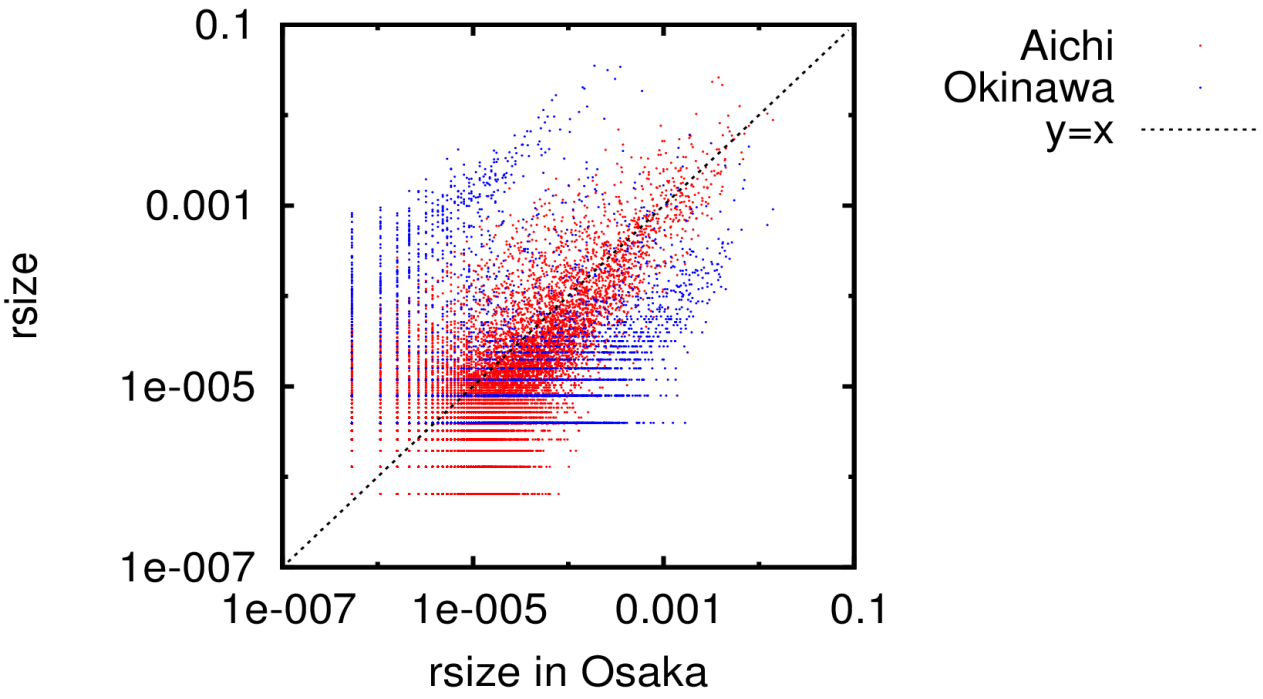


図 3.12 大阪と愛知の各姓の相対サイズ（赤点）、大阪と沖縄の各姓の相対サイズ（青点）の関係。 黒破線は y=x を表す。

この分布の内訳の類似性を定量化するために相対サイズのピアソンの相関係数R^{AB}を用いた。

$$\begin{cases} R^{AB} = \frac{T^{AB}}{T^A T^B} \\ T^A = \sqrt{\frac{1}{N}\sum (x_i^A - \frac{\sum x_i^A}{N})^2} \\ T^B = \sqrt{\frac{1}{N}\sum (x_i^B - \frac{\sum x_i^B}{N})^2} \\ T^{AB} = \frac{1}{N}\sum (x_i^A - \frac{\sum x_i^A}{N})(x_i^B - \frac{\sum x_i^B}{N}) \end{cases} \tag{3.1}$$

x_iは名前iの相対サイズs/S,Nは総姓数を示す。添え字 A,B はそれぞれ都道府県を示しており、T^A、T^Bは A,B それぞれにおける各姓の相対サイズの標準偏差、T^{AB}は A における各姓の相対サイズと B における各姓の相対サイズの共分散を示している。他方の県にしかない名前の相対サイズは 0 となる。つまりNは A,B の和集合における総姓数、総名数を示している。

これより 47 都道府県すべての組み合わせの姓の相対サイズの相関係数をもとめ、ヒートマップにしたものが図 3.14 である。図 3.13 はヒートマップの縦軸と横軸の値に対応した都道府県コードである。図 3.14 から各地方で相関係数が高いブロック構造をなしていることがわかる。このことから姓には地域性があるといえる。沖縄と他の県との相関係数はすべて低くなっており、沖縄特有の姓の相対サイズが大きいといえる。また同じように東北地方の県と他の県の相関係数は低くなっており、東北地方特有の姓の相対サイズが大きいといえる。

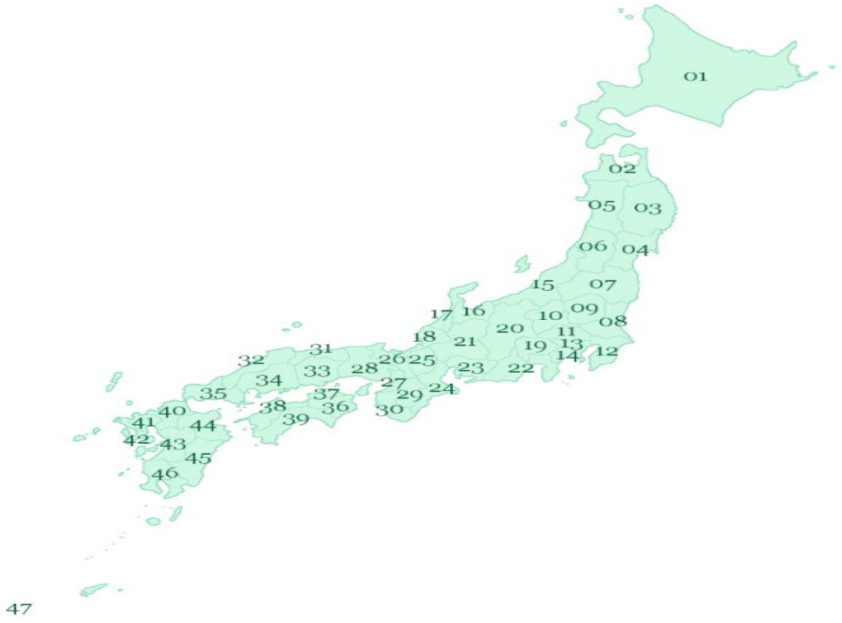


図 3.13 都道府県コード

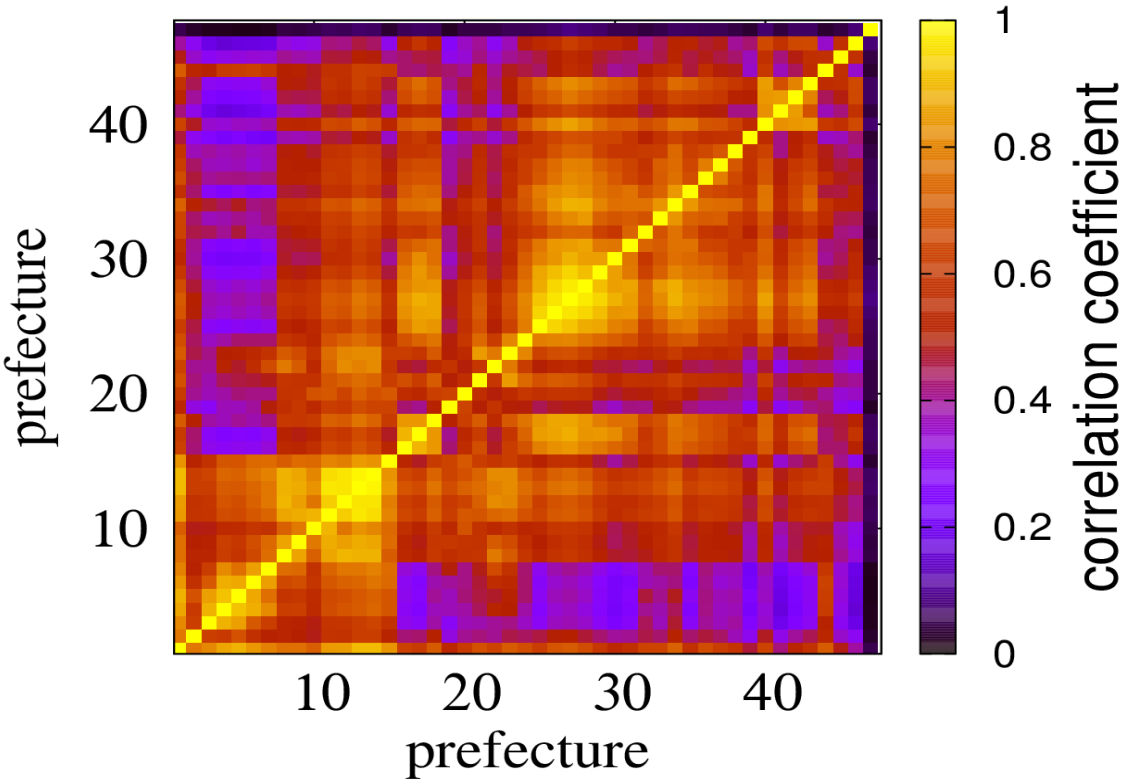


図 3.14 各都道府県での姓の相対サイズs/S（Sは総世帯数）の相関係数。縦軸と横軸はともに都道府県コード。

図 3.14 のようなヒートマップとは別に階層クラスター分析を行った。階層クラスター分析とは複数の要素間に非類似度を定義して、非類似度の小さな要素対からクラスタ化し、クラスタ同士も階層的にクラスタ化することで、要素間の関係を定量化する解析手法である。ここでは要素として都道府県を考え、非類似度は姓の分布関数から決める。具体的には、各都道府県を全ての姓の相対サイズで張られる空間内の点で表し、非類似度を都道府県を表す点間の距離で与える。姓の相対サイズの空間は 131,481 次元であり、名前の場合は 428,817 次元となる。この分析を行うために統計処理言語 R を用いた。その際使ったプログラムを付録 B に示す。対象間の距離としては、一般的に使われているユークリッド距離を用い、クラスター間の距離はウォード法と最遠隣法を用いたデンドログラムが図 3.16 である。ウォード法を用いた理由としてはクラスター同士の結合後のクラスターの分散と結合前のクラスターそれぞれの分散の和との差が最小になるクラスターのペアを併合するため、はずれ値に強いからである。最遠隣法を用いた理由は 2 つのクラスターの要素の全ての組み合わせの中で、最も近いのをクラスター間の距離とするため、はずれ値に弱く、はずれ値に強いウォード法と比較するためである。図 3.16 で色がついている都道府県は図 3.15 で同じ色の地域に属している。図 3.16 から、ヒートマップと同じようなブロック構造をみることができる。ヒートマップからは見づらい性質が 2 つある。1 つ目は、東北地方で地域性がみてとれるが青森だけはずれている。2 つ目は、北海道と関東地方が近い関係にあることがわかる。

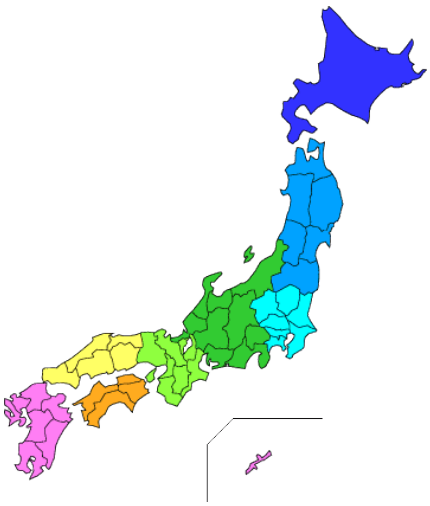


図 3.15　各都道府県を地域ごとに色分けしたもの。

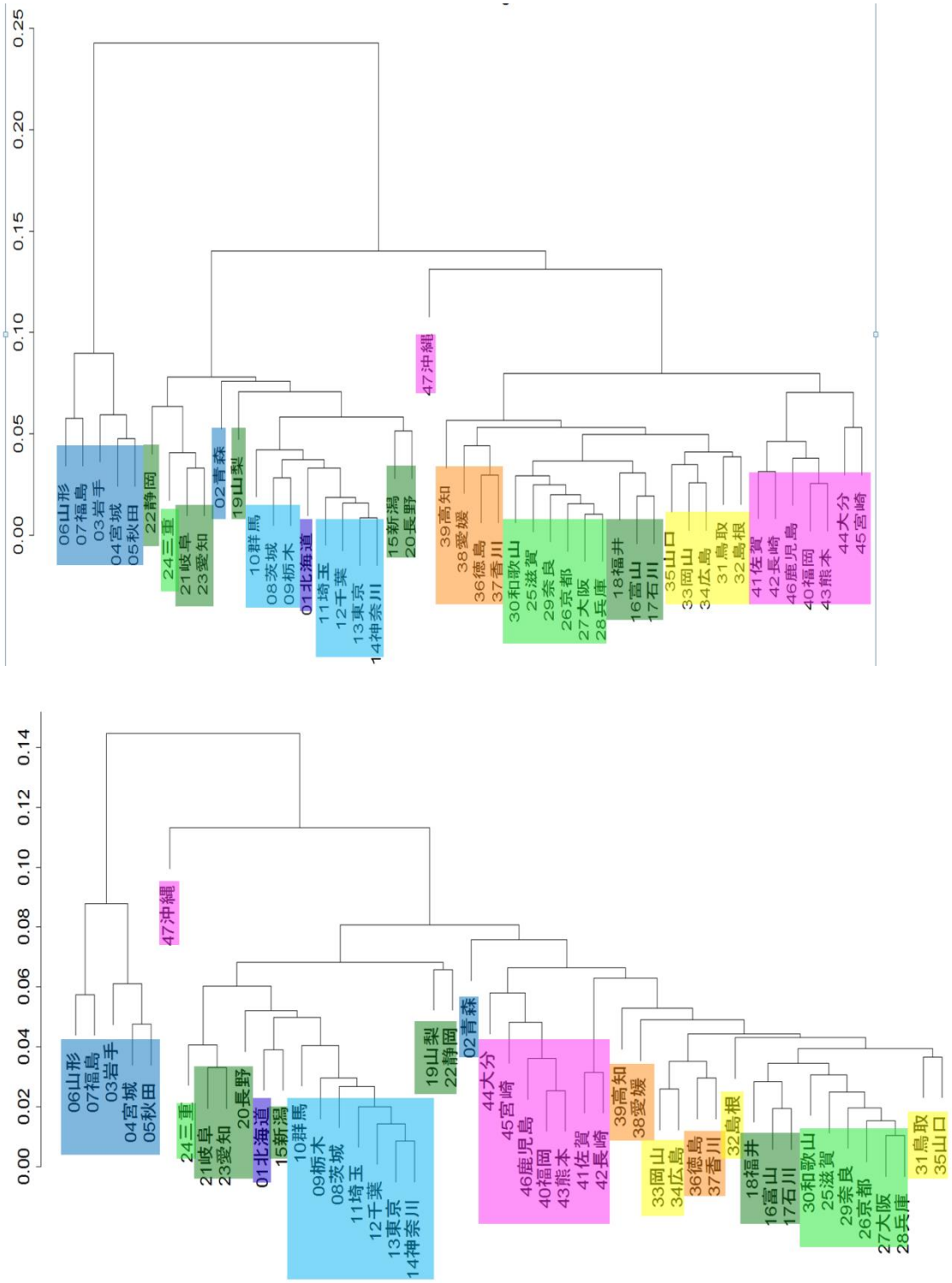


図 3.16 各都道府県の各姓の相対サイズをクラスター分析して作ったデンドログラム。距離はいずれもユークリッド距離。クラスタリング法はウォード法（上）と最遠隣法（下）。

各都道府県での相対サイズの相関係数の他に各都道府県での共通するランクの相関係数、他方にしかない姓のサイズ、他方にしかない姓の種類数、各都道府県での他方にしかない姓の相対サイズを図 3.17~図 3.20 で示す。相対サイズのヒートマップを作る際に他方の県にしかない姓のランクをつけづらいと述べていたが、図 3.17 は両県で共通する名前だけを抽出しそのランクの相関係数をヒートマップで表したものである。この図から、相対サイズのヒートマップと同じようにブロック構造を見ることができる。沖縄に関しても他の都道府県との相関係数が低く、東京(13)や大阪(27)などのような人口が比較的多い県との相関係数が低くなっている。図 3.18～図 3.20 は片方の都道府県にしかない統計量の図で、これまでのヒートマップには縦軸と横軸の交換に対する対称性があったが、これらにはそのような対称性がない。図 3.18 からは比較的人口が多い県にあって他の県にはない姓のサイズが大きく、その逆の姓のサイズが小さい事がわかる。これは、Heaps 則から人口と姓の種類数がベキ的な関係であることから人口が多い県は姓の種類数が多いためである。図 3.19 でも同様なことがいえる。図 3.20 は図 3.18 の縦軸と横軸をその県の電話帳の総世帯数で割ったもので、図 3.18 とは違って比較的人口が多い県にあって他の県にはない姓の相対サイズは特に大きくないが、その逆は小さい事がわかる。また、沖縄にしかない姓の相対サイズは特別大きい事がわかる。これは図 3.14 で述べた事と一致している。

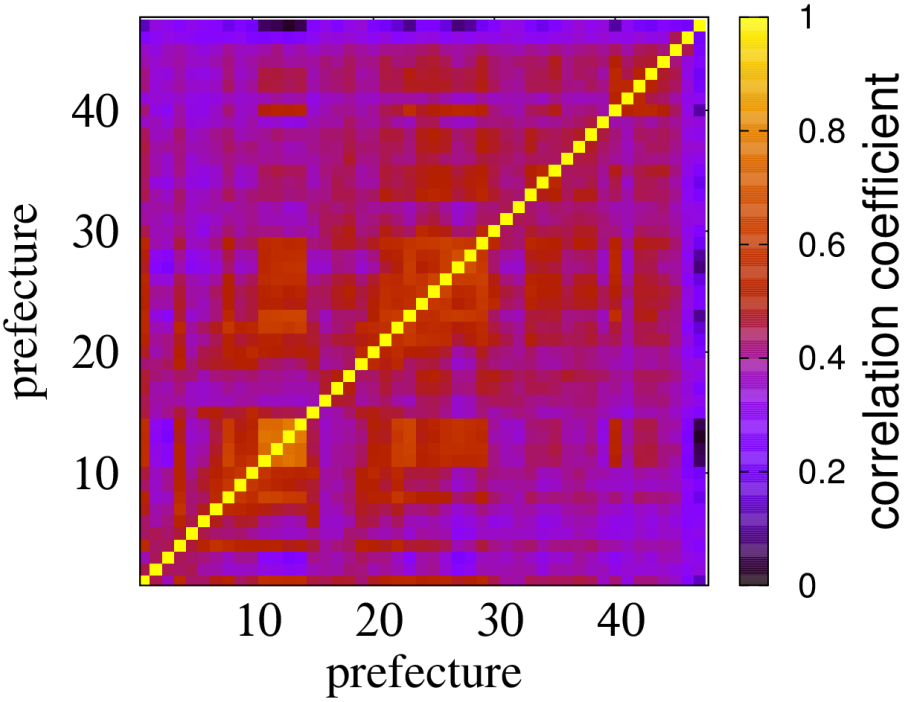


図 3.17 各都道府県での共通する姓のランクの相関係数。縦軸と横軸はともに都道府県コード。

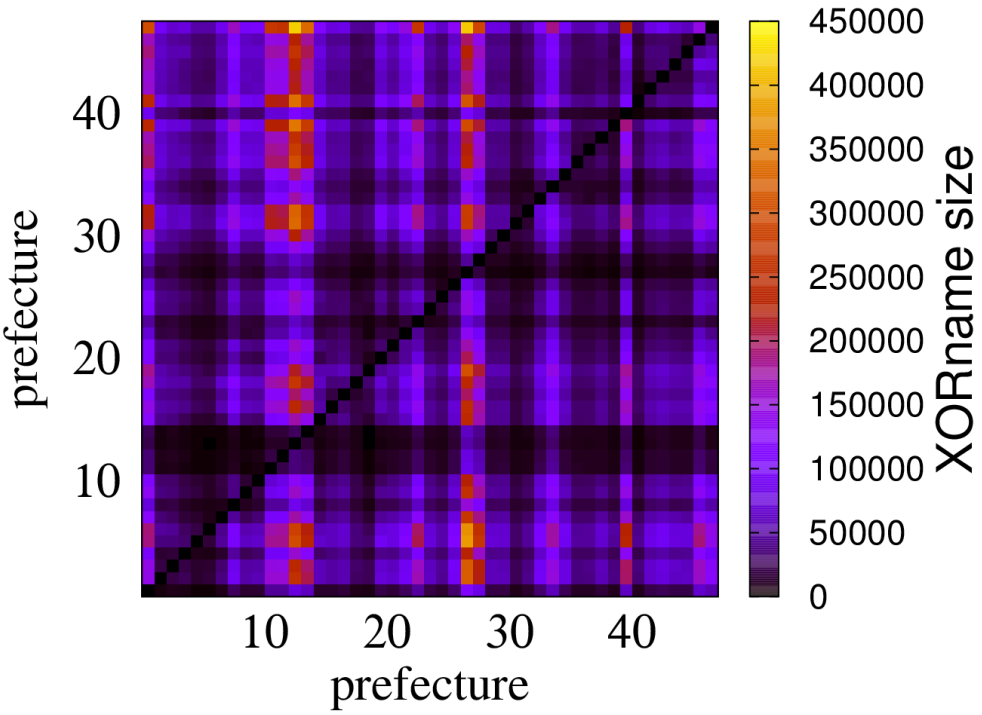


図 3.18　一方の都道府県にしかない姓のサイズ。横軸で示されたコードの都道府県にあり、縦軸の都道府県にない姓のサイズをプロットしたもの。縦軸と横軸はともに都道府県コード。

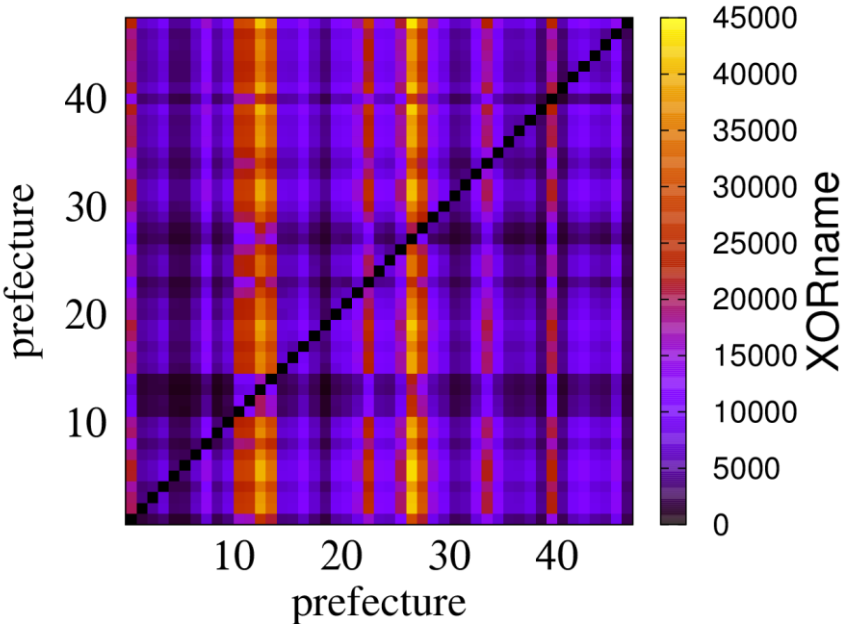


図 3.19　一方の都道府県にしかない姓の種類数。横軸で示されたコードの都道府県にあり、縦軸の都道府県にない姓の種類数をプロットしたもの。縦軸と横軸はともに都道府県コード。

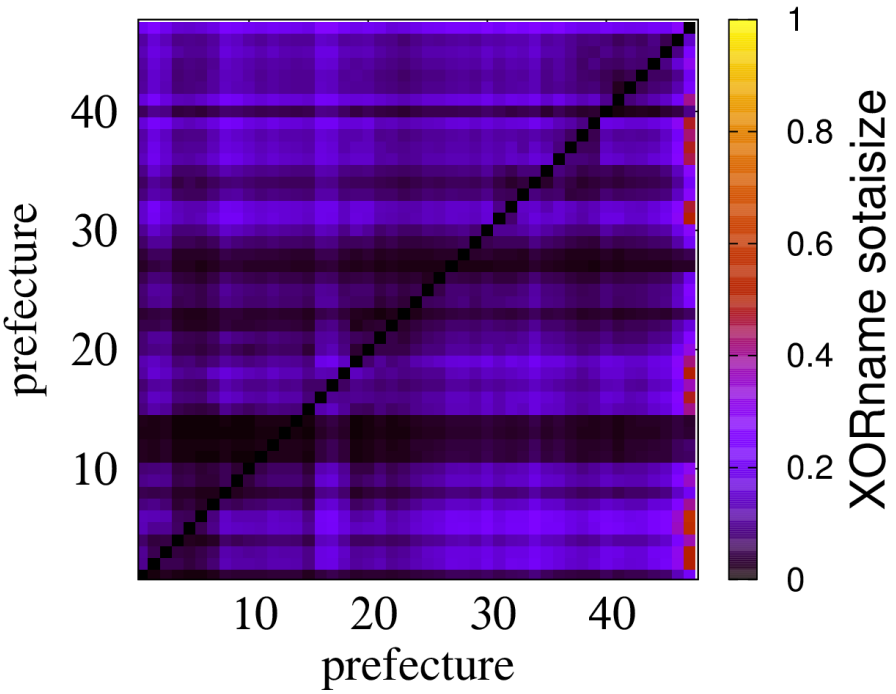


図 3.20　一方の都道府県にしかない姓の相対サイズ。横軸で示されたコードの都道府県にあり、縦軸の都道府県にない姓の相対サイズをプロットしたもの。縦軸と横軸はともに都道府県コード。

3.3.2 名についての分布の解析結果

名についても姓と同様に各都道府県での相対サイズの相関係数とデンドログラム、各都道府県での共通するランクの相関係数、一方にしかない名前のサイズ、一方にしかない名前の種類数、各都道府県での一方にしかない名前の相対サイズを調べた。図 3.21 で図 3.14 と比べると比較的に相関係数が全体的に高い事から分布の偏りが小さい事がわかる。ただし、電話帳データを用いている事から主として世帯主の名前が登録されており、比較的年配の男性の名前で登録されている件数が多いといったバイアスがかかっている。また、各地方で相関係数が高いブロック構造をなしており、名前にも地域性があることがいえる。沖縄についても姓と同様に相関係数が低い傾向にある。図 3.22 では(上)と(下)ともにブロック構造を見ることができる。青森は姓と違って東北地方と近い関係にあり、(上)では姓と同じく関東地方が近い関係にあるが、(下)では四国地方が近い関係になっている。クラスタリングの方法によってはこのような違いが生じるが、少なくとも北海道と東北地方は近い関係であるとはいえない。これより姓と名ともに地域性があるとはいえるが、距離が近ければ関係性が近いとはいえない事がわかる。図 3.23～図 3.26 に関しては姓の結果と同様なことがいえる。ただし、図 3.26 に関しては確かに沖縄にしかない名前の相対サイズは大きいが、姓と比べるとさほど高くない。

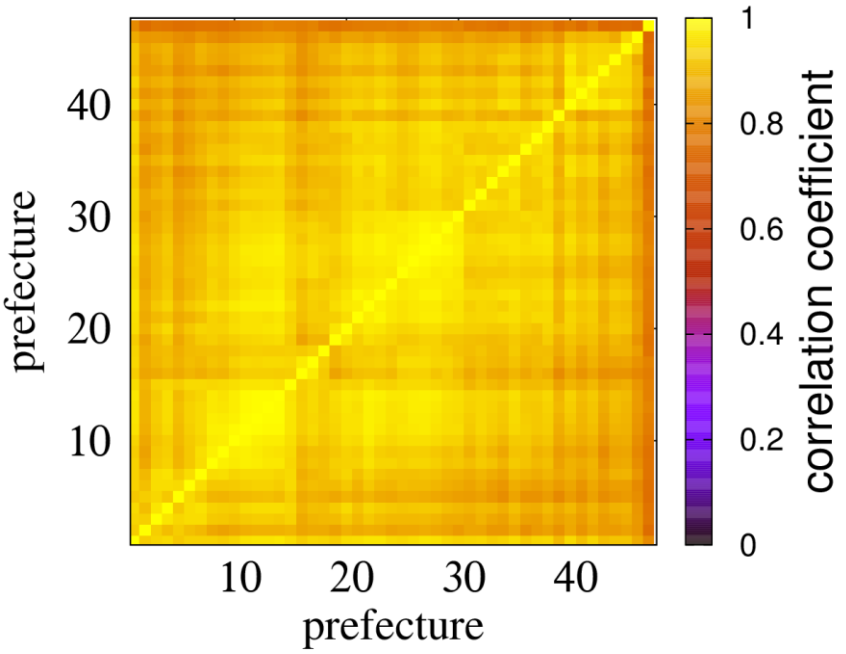


図 3.21 各都道府県での名前の相対サイズ（*s*/その都道府県の総世帯数）の相関係数。縦軸と横軸はともに都道府県コード。

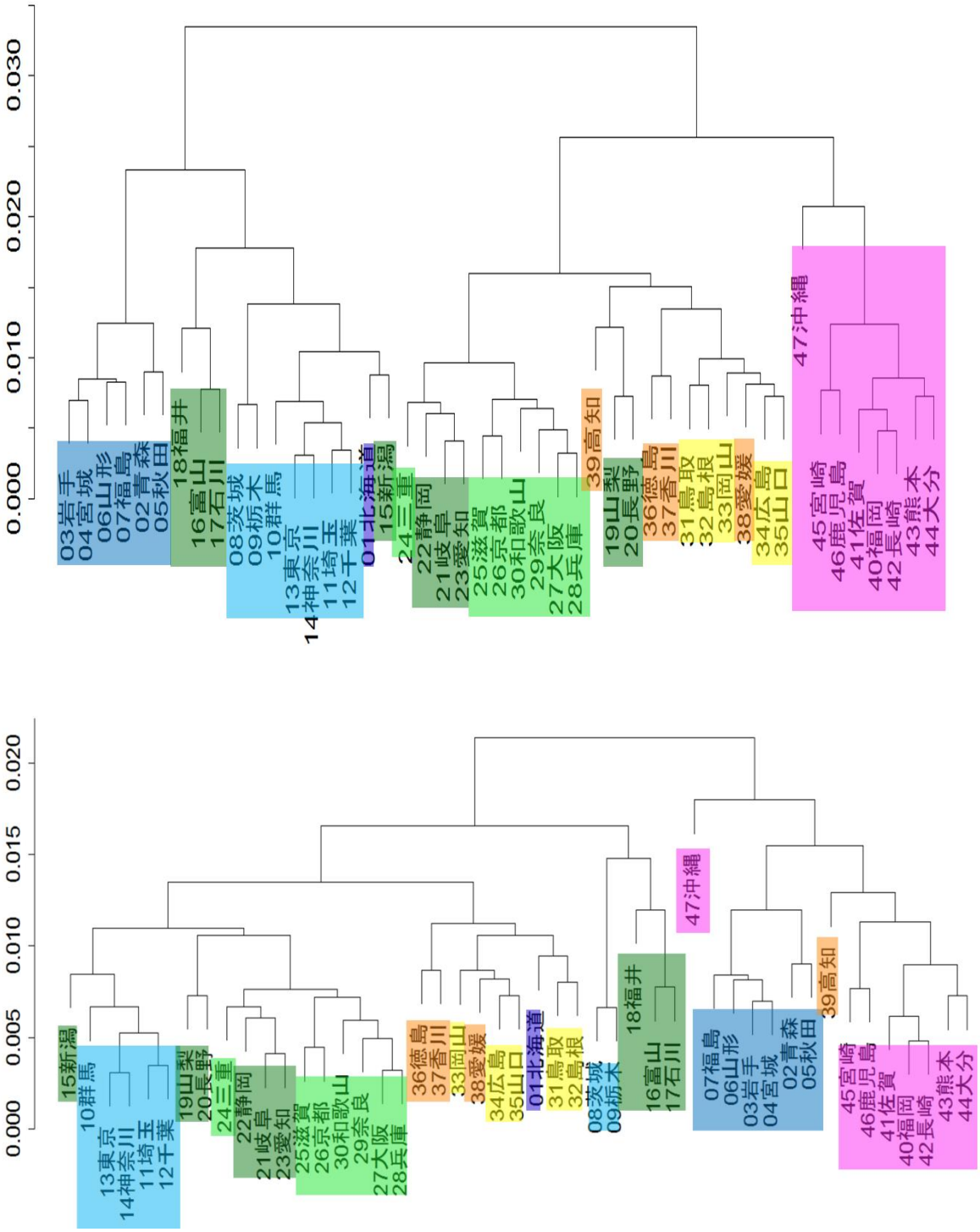


図 3.22　各都道府県の各名前の相対サイズをクラスター分析して作ったデンドログラム。距離はいずれもユークリッド距離。クラスタリング法はウォード法（上）と最遠隣法（下）。

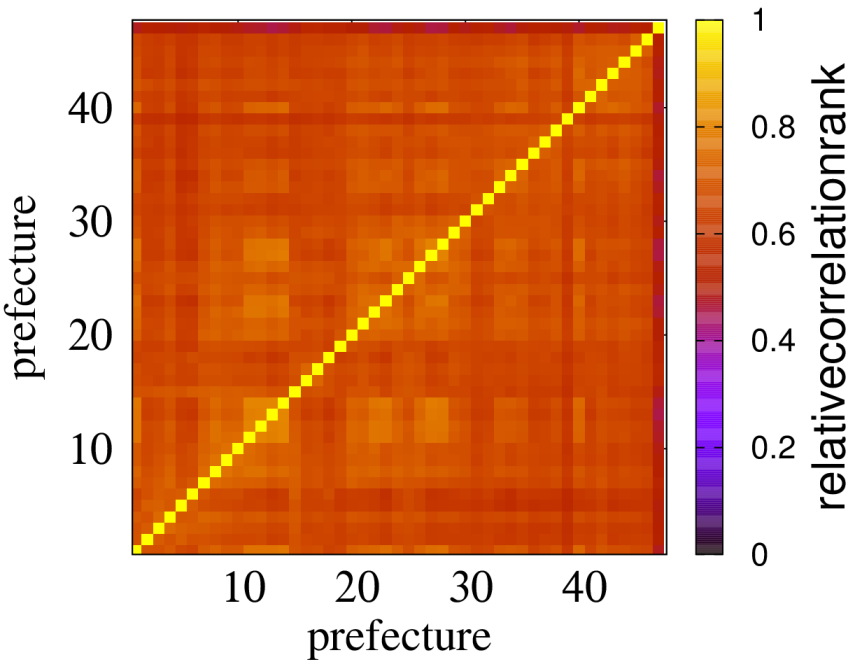


図 3.23 各都道府県での共通する名前のランクの相関係数。縦軸と横軸はともに都道府県コード。

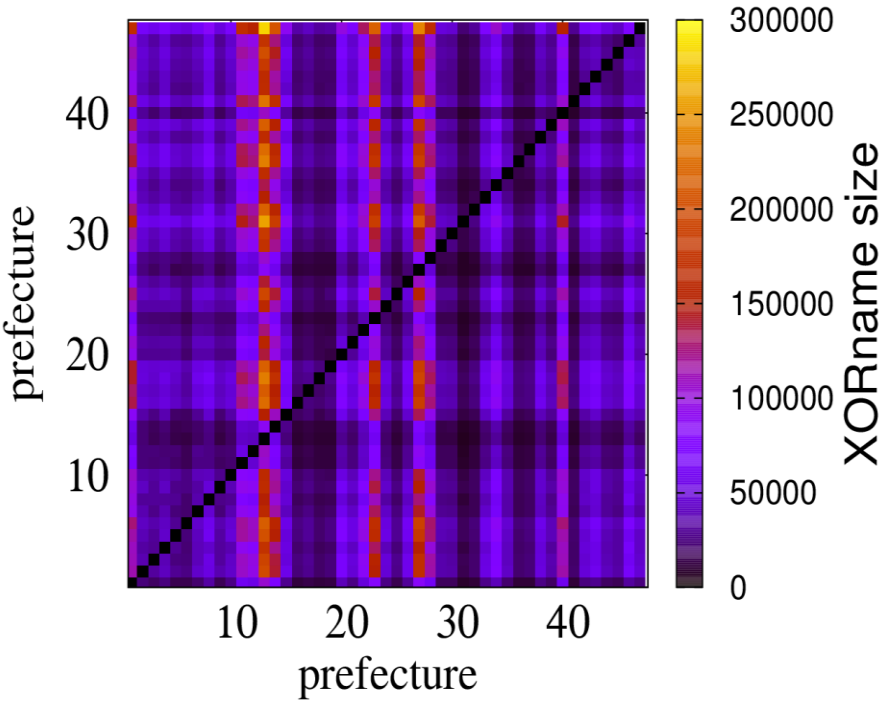


図 3.24 一方の都道府県にしかない名前のサイズ。横軸で示されたコードの都道府県にあり、縦軸の都道府県にない名前のサイズをプロットしたもの。縦軸と横軸はともに都道府県コード。

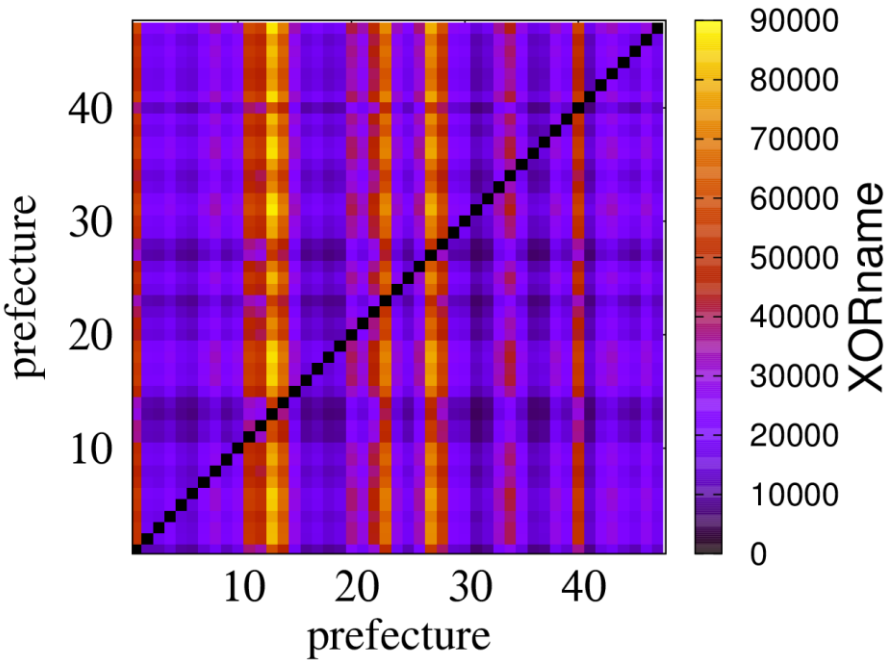


図 3.25　一方の都道府県にしかない名前の種類数。横軸で示されたコードの都道府県にあり、縦軸の都道府県にない姓のサイズをプロットしたもの。縦軸と横軸はともに都道府県コード。

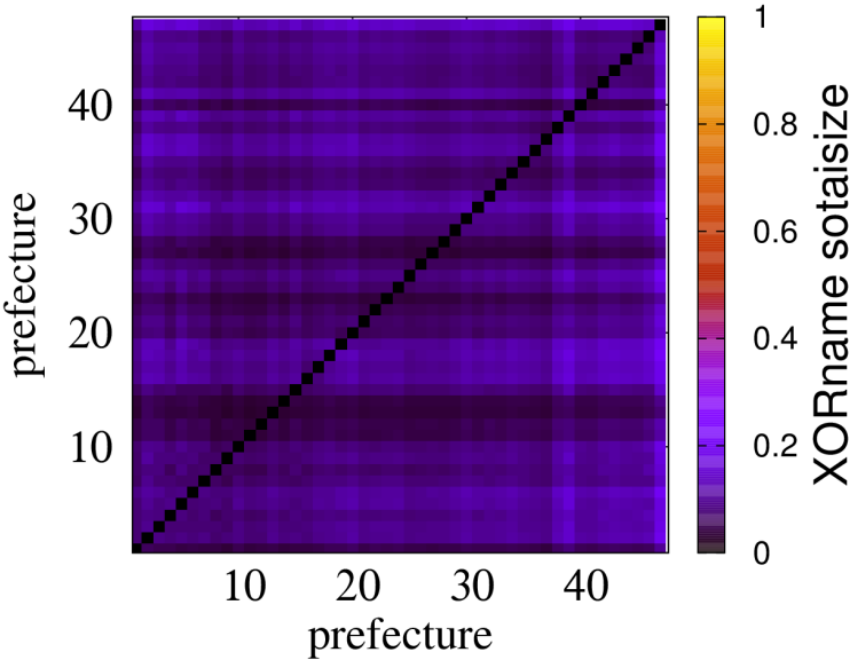


図 3.26　一方の都道府県にしかない名前の相対サイズ。横軸で示されたコードの都道府県にあり、縦軸の都道府県にない名前の相対サイズをプロットしたもの。縦軸と横軸はともに都道府県コード。

第 4 章　モデル

本章ではそれぞれの姓のサイズが時間とともにどう変化するかを記述するモデル方程式を提案する。

4.1 素過程

ある姓のサイズが増減する過程としては、出生、死亡、移動、婚姻などの様々なものがある。まず、それら一つ一つがどのようにモデル化されるか考えよう。なお、各姓のサイズは十分大きいと仮定し、それぞれの過程と姓の間には相関はないものと仮定している。

出生：多くの場合、姓は親のものを継承する。従って、単位時間（例えば一年）あたりの出生による姓のサイズの増加率はその年のその姓のサイズに出生率をかけたものである。

死亡：ある姓を持つ人が死亡するとその姓のサイズは 1 だけ減る。単位時間あたりの死亡による姓のサイズの減少率はその年のその姓のサイズに死亡率をかけたものである。

移動：ある姓を持つ人が都道府県 A から B に移動するとそれぞれの姓のサイズは 1 だけ増減する。この場合、単位時間あたりの移動による変化はその年のその姓のサイズに移動率をかけたものになる。

婚姻：ある姓を持つ人が婚姻した場合、多くの場合女性の姓が 1 減り、男性の姓が 1 増えると考えられる。故に婚姻する男女の姓によって結果は変わる。

ここでは単純化して、出生、死亡、移動の三つの過程だけを考える。

4.2 モデル方程式

簡単のため、国全体をある都道府県 A とそれ以外$\overline{\mathrm{A}}$に分ける。また、時間の単位を年とする。ある年 Y の都道府県 A で名前 j のサイズを$s_j^{A,Y}$とする。すると、

$s_j^{A,Y}$と$s_j^{\bar{A},Y}$が十分大きい時、$s_j^{A,Y}$、$s_j^{\bar{A},Y}$の発展方程式は以下のように表すことができる。

$$s_j^{A,Y+1} = s_j^{A,Y} + \left(\beta^{A,Y} - \delta^{A,Y} - \mu_{A\to\bar{A}}^{Y}\right)s_j^{A,Y} + \mu_{\bar{A}\to A}^{Y} s_j^{\bar{A},Y} \tag{4.1}$$

$$s_j^{\bar{A},Y+1} = s_j^{\bar{A},Y} + \left(\beta^{\bar{A},Y} - \delta^{\bar{A},Y} - \mu_{\bar{A}\to A}^{Y}\right)s_j^{\bar{A},Y} + \mu_{A\to\bar{A}}^{Y} s_j^{A,Y} \tag{4.2}$$

ここで、$\beta^{A,Y}$はある年 Y の都道府県 A の出生率、$\delta^{A,Y}$はある年 Y の都道府県 A の死亡率、$\mu_{\bar{A}\to A}^{Y}$はある年 Y の$\bar{A}$から都道府県 A への移動率である。これらのパラメータは名前 jにはよらないものと考えられる。

e-stat(https://www.e-stat.go.jp/)に各都道府県における人口、出生数、移動数についてのデータが公表されており、これらから出生率β、死亡率δ、移動率μを見積もることができる。

人口　：https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200524&bunya_l=02&tstat=000000090001&cycle=0&tclass1=000000090004&tclass2=000001051180&result_page=1&second2=1

出生数：https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00450011&tstat=000001028897&cycle=7&year=20170&month=0&tclass1=000001053058&tclass2=000001053061&tclass3=000001053064&result_back=1&second2=1

死亡数：https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00450011&tstat=000001028897&cycle=7&year=20170&month=0&tclass1=000001053058&tclass2=000001053061&tclass3=000001053065&result_back=1&second2=1

移動数：https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200523&tstat=000000070001&cycle=0&tclass1=000001051218&second2=1

パラメータに対してこれらの実測値を用いることで、初期条件が与えられれば方程式(4.1)、(4.2)を数値的に解くことができる。例えば、2002All（都道府県

別電話帳）から抽出した各都道府県の姓のサイズ（世帯数）に対して世帯あたりの平均人数及び電話帳への掲載率による係数を考慮することで Y=2002 における各姓の人口を見積もることができる。それを初期条件として用いて式(4.1)、(4.2)数値的に解き、Y=2010 での各姓の人口を推定する。最後に、再び電話帳に掲載されている世帯数に換算することで 2010 年の各都道府県における姓の世帯数を推定することが可能である。現段階では計算途中であるが、愛知県と北海道に対してなら 2010A、2010H から抽出した数値と直接比較できると期待される。婚姻の効果やサイズの小さな姓の揺らぎと絶滅過程、外国への移動などを考慮する必要性もあると考えられる。

また、モデル方程式(4.1)、(4.2)は日本全国を都道府県 A とそれ以外に分割したものであるが、47 都道府県に対する方程式に拡張することも可能である。その場合、全ての都道府県間の移動率$\mu^{Y}_{B\to A}$の組が必要となるが、移動数のホームページには都道府県間の移動数も好評されており、それを元に移動率も見積もることができる。

名のサイズの発展を記述するモデル方程式も同じように考えることができるだろうか。死亡と移動の過程は式(4.1)、(4.2)と同じ形で良いと考えられるが、生成は異なる。というのも、姓と異なり名は親のそれを受け継ぐことはないからである。名がどのように決まるかは未解明の問題であるが、例えば早川は Yule-Simon 過程を拡張したモデルを提案している[2]。それは、

(i) 一定の確率αでそれまでになかった新しい名前がつけられる。
(ii) それ以外（確率$1-\alpha$）では、それまでに存在した名前からその累積サイズに比例して選ばれる。
(iii) 親兄弟の名前は選ばれない。

という比較的単純なルールで決まるもので、適切な条件下で名前のランクサイズ分布がべき的になることも数値的に確かめられている。

もしこのモデルを採用するならば、生成過程は出生率βだけでなく、新名発生率αも含んだ式になると考えられる。3 章で述べた名前の地域性をどう取り入れるかなどの点も未解明であり興味深い問題であるといえよう。

第 5 章　結論

5.1 まとめ

電話帳（愛知県・北海道）からデータを取得し、姓・名の総世帯数 S に対する総姓数 N の分布である SN 分布とランクサイズ分布を統計的に解析した。SN 分布については姓・名いずれもべき則関係にあることがわかった。つまり Heaps の法則が成立していた。ランクサイズ分布については比較的ランクが大きい範囲に限るが姓・名いずれも総世帯数 S によることなくべき則の関係にあることがわかった。つまり Zipf の法則が成立していた。また総世帯数 S のべき乗でスケールする事により、ランクサイズのグラフが一つに重なることがわかった。

都道府県別の電話帳である 2002ALL を用いて姓と名前の分布を統計的に調べた。各都道府県の各姓の相対サイズを求め、その類似性を相関係数で表すヒートマップとデンドログラムで示した。各都道府県での姓と名前の相対サイズの相関係数は各地方で高い値をとっており、ブロック構造をなしていた。このことから姓だけでなく名前にも地域性があることがわかった。R を用いてデンドログラムを作成しそこからもヒートマップと同じくブロック構造をみてとれた。また、姓と名ともに地理的には近い北海道と東北地方が特に近い関係でないことがみてとれた。それぞれの姓のサイズが時間とともにどう変化するかを記述するモデル方程式を提案した。

5.2 今後の課題

今後の課題としては、全国、都道府県のスケールを都道府県、市町村のスケールに変えた際に非一様性がどうなっているかを確認する。名前に関しても地域性が見られたためその原因を探る。ある区域の特定の名前のサイズが時間とともにどう変化するかについて考える。

参考文献

[1] S. Miyazima, Y. Lee, T. Nagamine, and H. Miyazima, "Power-law distribution of family names in Japanese societies" , Physica A **278** (2000) 282-288.

[2]早川 良, "日本人の名前のサイズ頻度分布" , 平成 23 年度, 大阪府立大学大学院工学研究科修士論文.

[3] Seung Ki Baek, Hoang Anh Tuan Kiet, and Beom Jun Kim, " Family name distributions: Master equation approach" , Physical Review E **76** 046113 (2007) 1-7.

[4] Hoang Anh Tuan Kiet, Seung Ki Baek, and Beom Jun Kim, "Korean Family Name Distribution in the Past" , Journal of Korean Physical Society **51** (2007) 1-5.

[5] Damian H. Zanette and Susanna C. Manrubia, "Vertical transmission of culture and the distribution of family names" , Physica A **295** (2001) 1-8.

[6] Adrian Dragulescu, Victor M. Yakovenko, "Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States" , Physica A **299** (2001) 213-221.

[7] Luis M. A. Bettencourt, Jose Lobo, Dirk Helbing, Christian Kuhnert, and Geoffrey B. West, "Growth, innovation, scaling, and the pace of life in cities" , PNAS **104** (2007) 7301-7306.

[8] M.E.J. Newman, "Power laws, Pareto distributions and Zipf's law" , Contemporary Physics **46** (2005) 323-351.

謝辞

本研究に取り組むにあたり、様々な方々にお世話になりました。特に細部に渡りご指導いただきました水口毅准教授には、深く感謝致します。発表の練習の際には、大同寛明教授、堀田武彦教授、及川典子准教授、福田浩昭講師、先輩方、同回生、後輩の皆様には的確なアドバイスをいただき、ご支援いただきました。また両親には大学進学を支援していただきとても感謝しています。この感謝の念を忘れずにこれからも日々精進していきたいと思います。

付録

A. 全国と各都道府県における総世帯数、総姓数、総名数

表 A.1 全国と各都道府県における総世帯数、総姓数、総名数

	総世帯数	総姓数	総名数
全国	29,284,616	131,481	428,817
北海道	1,477,461	27,151	78,434
青森	383,313	8,908	38,109
岩手	365,882	9,779	37,650
宮城	492,673	12,848	42,986
秋田	345,761	6,792	34,797
山形	321,775	6,826	32,909
福島	497,778	11,017	43,842
茨城	710,180	19,312	51,256
栃木	453,725	10,000	40,076
群馬	492,456	12,989	45,477
埼玉	1,398,888	33,718	79,790
千葉	1,250,591	34,265	75,662
東京	2,103,225	45,868	109,154
神奈川	1,707,266	40,151	92,554
新潟	628,612	12,529	50,049
富山	290,852	12,246	32,829
石川	308,121	16,292	35,687
福井	215,804	10,906	30,615
山梨	239,689	7,960	33,441
長野	628,322	14,276	59,269
岐阜	531,842	15,311	50,969
静岡	911,358	20,912	68,236
愛知	1,548,115	31,801	94,440
三重	478,923	16,773	49,190
滋賀	324,877	16,200	37,676
京都	634,989	25,192	55,488

大阪	1,862,501	48,084	101,131
兵庫	1,290,566	37,650	81,360
奈良	360,436	20,138	41,915
和歌山	310,360	14,232	39,496
鳥取	165,393	8,678	27,165
島根	219,136	9,333	31,902
岡山	507,226	17,749	52,762
広島	715,924	26,330	62,578
山口	418,232	17,145	44,776
徳島	224,951	10,584	29,614
香川	265,581	10,651	31,558
愛媛	407,318	12,163	49,072
高知	247,325	7,439	39,756
福岡	1,052,739	27,778	77,158
佐賀	205,686	8,202	30,260
長崎	417,178	14,322	45,755
熊本	466,419	15,915	49,681
大分	328,278	12,917	41,438
宮崎	295,718	11,937	37,131
鹿児島	524,442	17,329	47,010
沖縄	252,900	6,302	33,143

B. デンドログラム描画のプログラム

デンドログラムを作成する際に用いた統計処理言語 R のコードは以下の通りである。

```
> my_data=data.frame(read.table(file="all.data",sep=",",header=F,row.names=1
))
> my_dist=dist(my_data,method=”euclidean”)
> my_hclust=hclust(d=my_dist,method=”ward.D”)
> plot(my_hclust)
```

all.data の中身は全国の名前とそれに対する各都道府県の相対サイズである。距

離とクラスタリングを変えるには R のコードの 3,4 行目の“　”を変えればよい。

C. スケールを変えた解析結果

3.3.1 節と 3.3.2 節では全国を 47 都道府県に分割したもので解析を行ったが、スケールを全国から県そして県を市区町村に分割したもので解析した結果を報告する。用いたデータは 3.2 節で取り扱っていた愛知(59 市区町村)と北海道(105 市区町村)のデータである。ヒートマップからは相関係数の高低がみられるが、その縦軸と横軸は市区町村の名前順に並んでいるので、構造が掴みにくい。一方、デンドログラムを見ると名古屋市が一つのグラスタを形成していることがわかる。

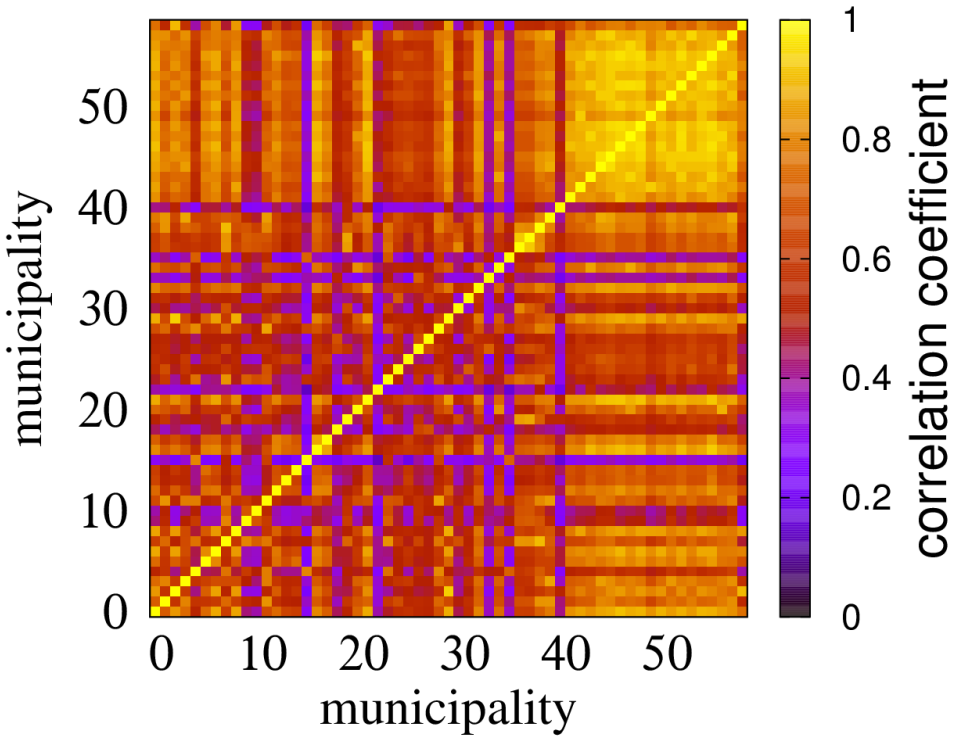


図 C.1 愛知県内の市区町村での姓の相対サイズ s/S (Sは総世帯数) の相関係数。

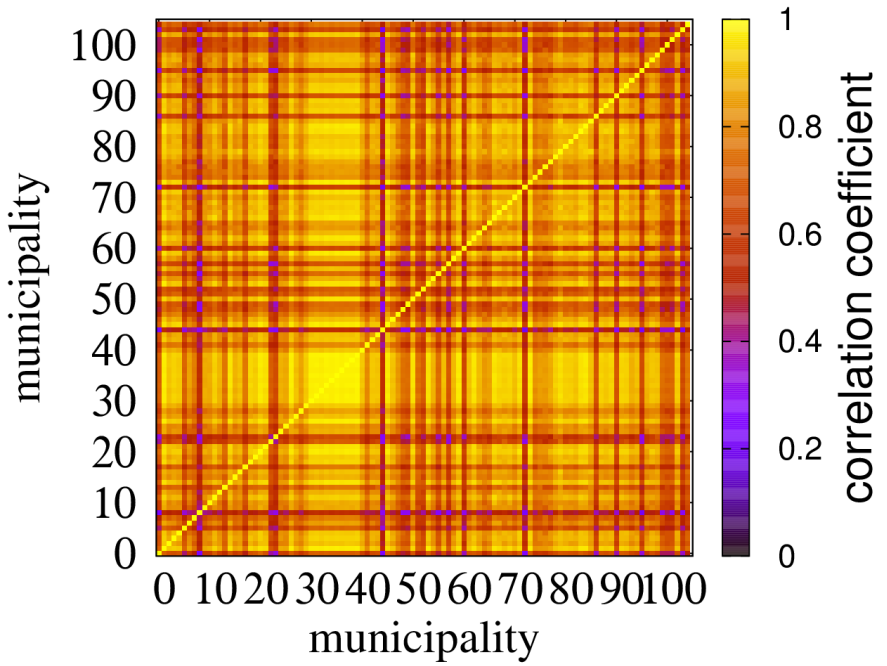
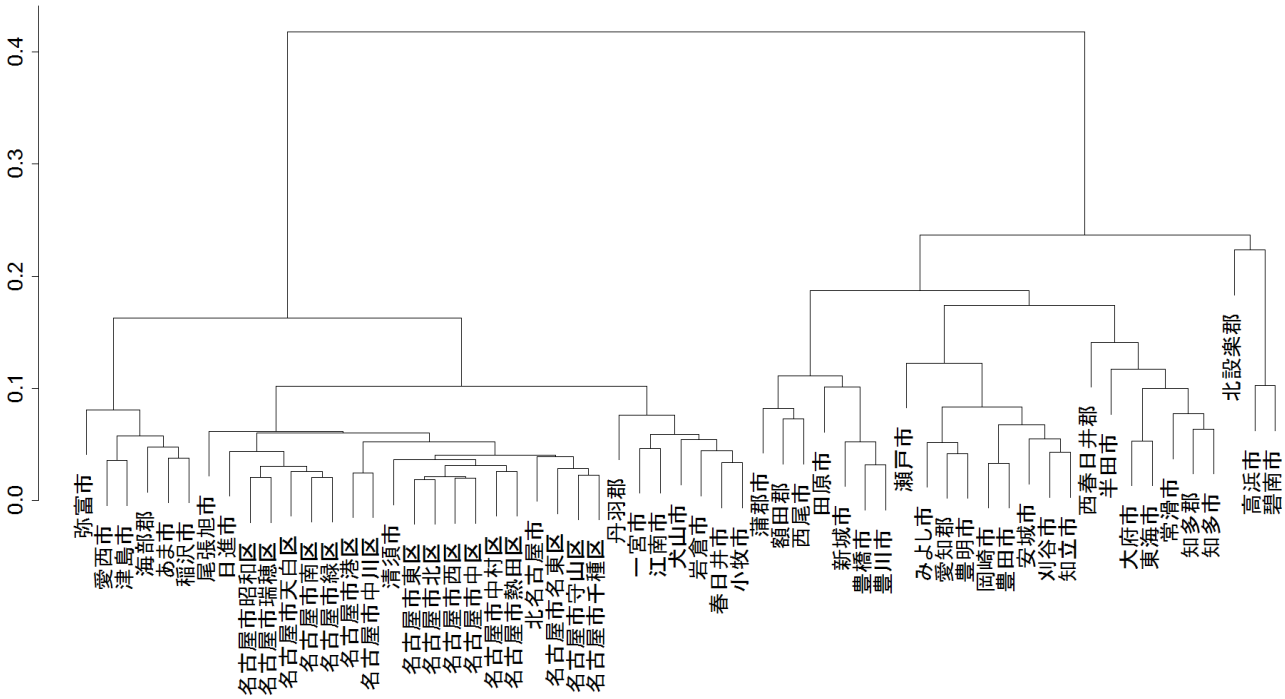


図 C.2 北海道内の市区町村での姓の相対サイズ s/S (Sは総世帯数) の相関係数。



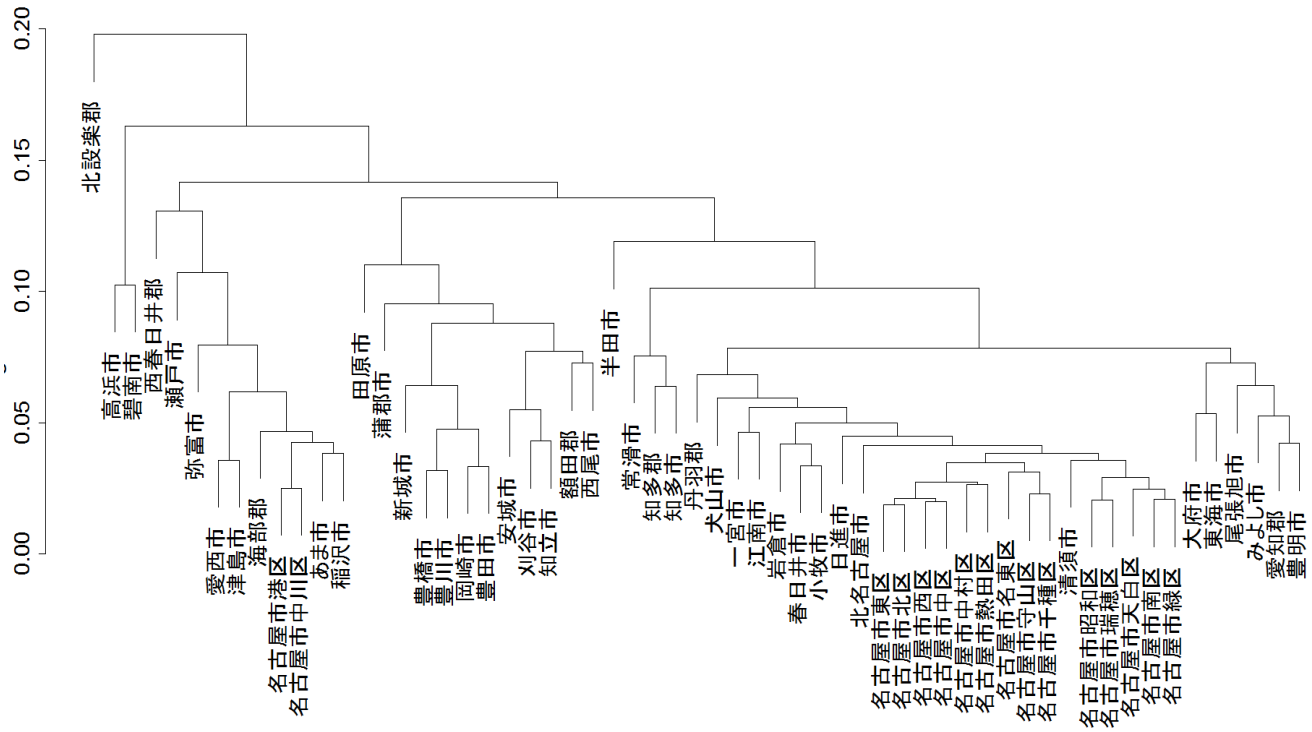


図 C.3　愛知の各姓の相対サイズをクラスター分析し表したデンドログラム。距離はいずれもユークリッド距離。クラスタリング法はウォード法（上）と最遠隣法（下）。