フィンランド 2002 年ビジネスレジスターの補完実験*

宮内　環（慶應義塾大学経済学部）†

概 要

本稿では、1990 年代後半よりマイクロデータにおける欠損値の補完に広く適用されるようになった NIM (Nearest-neighbor Imputation Method) による補完の精度の評価を、フィンランド国家統計局の 2002 年ビジネスレジスターを用いて行い、その結果について報告する。

当該ビジネスレジスターの特徴として次の二点が挙げられる。第一点目は、ビジネスレジスターを構成する事業所ファイルの事業所レコードには、各事業所の緯度・経度の情報が含まれており、任意の二つの事業所間の地理的距離が計算可能である、という点。第二点目は、このビジネスレジスターは事業所ファイル、企業ファイルの二種類により構成され、事業所レコードと企業レコードには各々識別番号が付与されており、事業所レコードと当該事業所が属する企業の企業レコードとを接続することが可能である、という点である。第一点目の事業所の緯度・経度情報は、事業所レコードのある一定の調査項目について生じる欠損値を補完するにあたり、NIM を適用するために不可欠である。さらに二点目により、事業所レコードのある一定の調査項目について生じる欠損値を NIM により補完する場合、当該事業所と同一の属性を持ち、当該の調査項目に欠損を生じていないドナーの候補としての事業所の範囲を特定するにあたり、事業所レコードに含まれる産業分類の情報だけを用いる場合、さらに欠損値が生じている当該事業所が属する企業の売上高階層の情報を併せて用いる場合、といった、情報量の差による補完の精度の比較が可能となる。

本稿では、フィンランド国家統計局のこうした特長を活かし、シミュレーションの手法により事業所の従業者数に擬似的に欠損値を生じさせ、これを NIM により補完を試み、その補完により得た値と、欠損を生じさせる前の値との比較により、補完の精度を評価した。

本稿における最も重要な結論は、つぎの二点である。第一点目は、ドナーの事業所の範囲を特定するにあたり、欠損値が生じた事業所に関する情報だけでなく、当該事業所が属する企業に関する情報も併せて用いると、補完の精度がより向上することが示される、という点。第二点目は、事業所の従業者数の欠損値を補完するにあたり、当該事業所の従業者階層に関する情報を併用すると、補完の精度が飛躍的に向上する、という点である。

キーワード:ビジネスレジスター、欠損値、補完、NIM(Nearest-neighbor Imputation Method)、シミュレーション
JEL 分類コード：C81

*この研究成果は科学研究費助成事業 (学術研究助成基金助成金 (基礎研究 (C)(一般))「ビジネスレジスターによる企業動態統計の開発」(補助事業期間　平成２４年度～平成２６年度) の助成を得ることにより具体化されたものである。この研究に用いたビジネスレジスターはフィンランド国家統計局より有償で譲り受けたもので、この取得のための費用は前述の助成金の一部により賄われた。さらに同科学研究費助成事業の助成により組織された研究会の研究代表者である菅幹雄 (法政大学経済学部)、同研究会の研究協力者である森博美 (法政大学経済学部)、宮川幸三 (慶應義塾大学産業研究所)、Jukka Pakola(Statistics Finland)、Ossi Nurmi(Statistics Finland) の各氏からは、研究会において当該研究成果に関する研究上の多大なる助言と示唆を得た。ここに記して感謝の意を表する。なお、本稿におけるすべての誤謬は筆者の責に帰するものである。

†email:miyauchi@econ.keio.ac.jp

1 はじめに

統計調査においては、たとえそれがセンサス調査であっても、ある一定の調査項目について欠損値が生じやすい、という傾向はしばしば経験されるところである。ある調査項目に欠損値が生じた場合、これを集計する前の個票の段階で補完を試みる方法が近年において盛んに議論され、その方法の有力なものとして Nearest-neighbor Imputation Method (以下では NIM とよぶ) が近年注目されている。

本稿では、この NIM における補完の精度、およびその補完における追加的な情報が補完の精度にいかなる効果を及ぼすか、の二点について客観的に評価するために、フィンランド国家統計局の 2002 年ビジネスレジスターのデータセットを用いて、シミュレーションの手法による補完に関する数値的な実験を行うことにした。本稿における実験では、主に次の二点について確認を行った。まず、欠損確率が一様である場合と個体の属性により変わる場合の各々において、補完の精度はどのように変化するか数値的に調べた。つぎに、補完に際して追加的に利用可能な情報が、補完の精度にいかなる与える影響を与えるかを数値的に調べた。

NIM による補完の精度を確認するにあたり、本稿の実験で用いるフィンランドの 2002 年ビジネスレジスターのデータセットの特徴として次の二点を指摘しておくべきであろう。

第一点目として、大別して次に述べる二種類のファイルより構成されていることである。その一つ目は、フィンランド国内の事業所ファイル、二つ目はフィンランド国内の企業ファイルである。前者の事業所データセットに含まれる各事業所レコードには、後者の企業データセットの各企業レコードと接続可能な識別番号が付与されており、これによって各事業所レコードを当該事業所が属する企業の企業レコードと接続することが可能となっている。

第二点目として、前者の事業所ファイルの事業所レコードには、事業所の従業者数や操業開始年月などの属性の他に、当該事業所の緯度・経度の情報が含まれているので、任意の二つの事業所間の地理的な距離を知ることができる。NIM による補完にはこの地理的な位置情報が欠かせない。

本稿における実験においては、シミュレーションの手法により事業所レコードに含まれる従業者数の数値項目に擬似的に欠損値を生じさせ、NIM により補完した結果と、当該数値項目に欠損値を生じさせる前の値との比較により、補完の精度に関して客観的な評価が可能となる。本稿の実験のデザインは主に次の二点である。

第一点目は、擬似的な欠損を与える確率を、事業所の属性とはかかわりなく一様とする場合、事業所の属性により変化する場合、の二通りを設定したことである。本稿の実験では、こうした欠損確率の分布の違いにより補完の精度がどのような影響を受けるのかを評価した。

第二点目は、欠損値が生じた事業所において観察される属性 (欠損値が生じている数値項目以外の属性) と類似の属性を持つ別の事業所を探すにあたり、観察される属性を、事業所レコードから得られる情報に限定した場合、事業所レコードだけでなく当該事業所が属する企業の企業レコードから得られる情報も併せて用いる場合、の二通りを設定したことである。すなわち、実験を行うにあたり、事業所レコードと当該事業所が属する企業の企業レコードを接続データセット、言わば事業所と企業の「名寄せ済みファイル」を後者の場合のために、事業所ファイルとは別に準備した。本稿における実験では、欠損値を NIM により補完する試みにおいて、ある調査項目に欠損値が生じている事業所と同一の産業分類に属し、当該項目に欠損値が生じていない他の事業所をドナーの候補として特定してい

る。このとき、欠損値が生じている事業所についてその産業分類以外の属性として、当該事業所が属している企業の売上高階層がわかれば、この情報を用いてドナーの候補をさらに限定することができよう。本稿における実験のデザインの第二点目は、事業所レコードより得られる産業分類の情報のみによりドナーの候補を特定した場合、「名寄せ済みファイル」により産業分類の他に、当該事業所が属する企業の売上高階層の情報も併せてドナーの候補を特定した場合、の各々において補完を行い、これら両者の場合の補完の精度を比較することを行った。

本稿の構成は次のとおりである。第2節では、先行研究の概略を述べ、第3節では、本稿の実験で用いたフィンランド国家統計局のビジネスレジスターのデータの概略について述べる。第4節では、本稿の実験方法について述べ、第5節では、実験の結果についてその概略を述べる。第6節では結語を述べる。

2 先行研究

欠損値の補完は、かねてより統計調査の課題であった。当初は集計表における欠損セルを補完する方法が主であったが、1970年代ころからマイクロデータが統計の活用において主流となり始めたとともに、マイクロデータの各レコードに含まれる欠損値の補完が主要な問題として議論されるようになった。Fellegi and Holt (1976) がマイクロデータの調査項目の観測値について、整合性の検討 (Editing) と異常値や欠損値を補完 (Imputing) をコンピュータを用いて行う方法を提示し、以後 "Fellegi-Holt Method" として知られている。さらに同時期に Rubin (1976) も欠損値についての推測の議論を行っている。Little and Rubin (1987;2002) も含め、これらは主に統計的方法を背景としている。その後、Winkler and Chen (2001) では、"Fellegi-Holt Method" の展開が示されている。その他に、Rubin (1987;1996) は単一の欠損値だけでなく、複数の欠損値のセットを同時に補完する方法について提示している。すなわち欠損値の個々についてその周辺分布に基づいて補完をするのではなく、複数の欠損値のセットの背後にある同時分布を想定して補完を行うという考え方である。

一方、Bankier, et al. (1997) の提案による NIM (Nearest-neighbor Imputation Method) をカナダの国勢調査へ適用し、その後 Statistics Canada (1999;2002) ではその展開を行い、以後マイクロデータにおける補完にこの NIM が適用される事例が増えている。Andridge and Little (2010) は、以上の補完に関する歴史的展開を踏まえ、補完の方法論に関する包括的な議論を行っている。

わが国では森・菅 (2012) が事業所企業統計をビジネスレジスターと位置づけ、その個票データを用いて補定の精度をシミュレーションにより検証している。そのシミュレーション手法の概略をおおよそ次の通りである。まず事業所企業統計調査の本調査である平成18(西暦2006)年のデータセットにおいて、従業者数の変数についてランダムに擬似的欠損値を作り出す。つぎに当該欠損値が生じた事業所が (1) 既存事業所であるか、あるいは (2) 新規事業所であるかにより、次の二通りの異なる補定方法を採用している。第1に当該欠損値が生じた事業所が既存事業所である場合には、過去に行われた平成16(西暦2004)年あるいは平成13(西暦2001)年調査における同一事業所の観測値を用いて補完する。第2に当該欠損値が生じた事業所が新規事業所である場合には、同じ平成18年の新規事業所のうち、欠損値を生じていない事業所における従業者数の地域別・産業別平均値により補完している。以上の方法により補完された従業者数と、実際 (擬似的に欠損値とされる前) の従業者

数の誤差 (および誤差率) の分布により、補完の精度を確認している。なお、欠損確率は、地域区分・産業区分にかかわりなく一様の場合、都道府県や市町村といった地域区分および産業区分によって変わる場合を試みている。シミュレーションの結果として、筆者は次の二点を挙げている。第一に、既存事業所の欠損値を過去のデータにより補完する方法は時系列的なトレンドが急な変化でなければ良好な結果を与えるであろう。第二に、新規事業所の欠損値を欠損値を生じていない事業所の観測値で補完する方法は、事業所の特性に関する情報 (たとえば地域区分、産業区分、資本金などの規模区分) の情報を活用できれば良好な結果を与えるであろう、としている。

わが国では他に、高橋・伊藤 (2013) は売上高の補完についての検討を行っている。

3 フィンランド国家統計局のビジネスレジスター

我々はフィンランド国家統計局から 2002 年のビジネスレジスター (以下では "FBR2002" とよぶ) を有償で譲り受け、これを使う許可を得た。ここでは、FBR2002 の概略と本稿で述べる NIM (Nearest-neighbor Imputation Method) による補完の実験のために準備されたデータセットの概略を述べる。

3.1 事業所ファイルと企業ファイル

FBR2002 は、事業所ファイルと企業ファイルの二つのファイルより構成されている。前者の事業所レコードと後者の企業レコードにはユニークな企業 ID が付与され、この企業 ID によってある企業を構成する事業所のレコードを名寄せすることが可能となっている。表 1 と表 2 には各々、事業所レコード、企業レコードに含まれる変数を掲載した。

表 1: FBR2002 の事業所ファイルの事業所 (レコード) 数と変数

事業所数	255,127	
変数	桁数	備考
企業 ID	11	FBR2002 の企業レコードとのリンケージに利用可能
事業所コード	9	
郵便番号	5	
行政区番号	3	
産業分類記号	5	
事業所の従業者数階層	1	0-4, 5-9, 10-19, 20-49, 50-99, 100-199, 200- の 7 階層
操業開始年月日	8	西暦 4 桁、月 2 桁、日 2 桁
事業所の従業者数	6	補完の実験ではこの変数を擬似的に欠損値とした。
事業所の位置座標	14	前 7 桁が緯度、後 7 桁が経度

なお、以上の事業所レコードおよび企業レコードの「従業者数」には欠損値のコードは見当たらない[1] 。従って、第 4 節で述べる補完の実験において、擬似的に「従業者数」に欠損値が生じたとしてこれを補完した場合、その補完の精度を、欠損を生じさせた当該レ

[1] 数値の零 (ゼロ) は頻出するが、筆者はこれを「従業者数」の欠損値とはみなしておらず、当該変数の観測値として 0 の値が得られたと解釈している。これに対し、「事業所の位置座標」には座標の値が代入されていないレコードが多数あり、この場合は明らかに欠損値であると判断できる。この点を考慮すると、「従業者数」に現れる 0 の値は欠損値ではなく、0 という観測値が得られたと理解することが妥当であろう。

表 2: FBR2002 の企業ファイルの企業 (レコード) 数と変数

企業数	226,426	
変数	桁数	備考
企業 ID	11	FBR2002 の事業所レコードとのリンケージに利用可能
行政区番号	3	
産業分類記号	5	
企業の従業者数	6	
企業の従業者数階層	1	0-4, 5-9, 10-19, 20-49, 50-99, 100-199, 200-499, 500-999, 1000- の 9 階層
売上高階層	1	9 階層
事業開始年月日	8	西暦 4 桁、月 2 桁、日 2 桁
法律上の組織形態区分	2	コードの詳細は不明
所有形態区分	1	私有国内, 国有, 地方自治体, オーランド自治州, 外国人, その他, の 6 区分
雇用する者の活動状況区分	1	コードの詳細は不明
付加価値税の状況区分	1	コードの詳細は不明
輸入・輸出の状況区分	1	輸入・輸出の有無の別

コードの「従業者数」の値と、補完によって得た値との乖離によって評価することが可能である。

本稿の第 4 節で述べる補完の実験では、上述の事業所ファイルのみを用いた実験の他に、FBR2002 の事業所レコードを、FBR2002 の企業ファイルの企業レコードに名寄せをしたレコード (以下では「名寄せ済みレコード」とよび、この名寄せ済みレコードから構成されるファイルを「名寄せ済みファイル」とよぶ) も用いた実験を行なっている。次項ではこの名寄せ済みレコード作成と名寄せ結果の概略について述べる。

3.2 事業所レコードの企業レコードへの名寄せ

「名寄せ済みレコード」の作製には、表 1 に示した事業所レコードの企業 ID と、表 2 に示した企業レコードの企業 ID とを比較し、両者が完全に一致した場合に、事業所レコードの企業レコードへの名寄せを行なった。名寄せの結果は非常に良好で、その状況を表 3 および表 4 に掲示した。

表 3: FBR2002 の事業所レコードの企業レコードへの名寄せ状況

場合	事業所レコード数
企業レコードに名寄せ（接続）できなかった事業所レコード数	760
企業レコードに名寄せ（接続）できた事業所レコード数	254,367
合計	255,127

表 4: FBR2002 の企業レコードの事業所レコードへの名寄せ状況

一つの企業レコードに名寄せ (接続) できた事業所レコード数	企業 レコード数
0	41
1	219,006
2	4,445
3	1,244
4	503
5	276
6	208
7	124
8	62
9	74
10	52
11	37
12	40
13	21
14	14
15	22
16	14
17	16
18	16
19	9
20	14
21- 50	125
51-100	38
101-	25
合計	226,426

4 補完の実験の概略

本稿で述べる補完の実験は、ある事業所レコードの従業者数について擬似的に欠損値が発生したとして、その欠損値を NIM により補完し、この補完によって得られた従業者数の値、および当該レコードに記録された従業者数を真の値とし、両者を比較して補完の精度を評価している。なお、欠損値の発生確率は、事業所の属性にかかわりなく一様に与えた場合、事業所の産業大分類および従業者階層別に異なる値を与えた場合を設定している。前者の場合は、一様に 15%, 30%, 45% の 3 ケースとし、これらを各々「ケース 1」、「ケース 2」、「ケース 3」とよぶ。さらに後者の場合は、事業所の産業大分類および従業者数規模別の欠損確率を表 5 に示し、これを「ケース 4」とよぶ。なお、産業大分類は大文字のアルファベット一文字で表され、その意味を表 6 に示した。

4.1 事業所ファイル、および名寄せ済みファイルによる 2 種類の補完の実験

本稿で述べる補完の実験は、用いたファイルによって 2 種類の実験に大別することができる。第 1 番目は、事業所ファイルを用いる実験であり、第 2 番目は名寄せ済みファイルを用いる実験である。補完のための NIM を用いる場合、欠損値が生じた事業所と類似の属性を持つ (欠損値が生じていない) 事業所を「ドナー」として探す必要があるが、このと

表 5: 産業大分類別・事業所従業員規模別に設定した欠損確率：産業・規模により異なる場合

産業	事業所従業員規模						
分類	0〜4	5〜9	10〜19	20〜49	50〜99	100〜199	200〜
A	0.30	0.25	0.25	0.15	0.15	0.05	0.05
B	0.30	0.25	0.25	0.15	0.15	0.05	0.05
C	0.30	0.25	0.25	0.15	0.15	0.05	0.05
D	0.15	0.15	0.15	0.10	0.10	0.05	0.05
E	0.15	0.15	0.15	0.10	0.10	0.05	0.05
F	0.30	0.25	0.20	0.10	0.10	0.05	0.05
G	0.30	0.25	0.20	0.10	0.10	0.05	0.05
H	0.30	0.25	0.20	0.10	0.10	0.05	0.05
I	0.30	0.25	0.20	0.10	0.10	0.05	0.05
J	0.15	0.15	0.15	0.10	0.10	0.05	0.05
K	0.30	0.25	0.20	0.10	0.10	0.05	0.05
L	0.15	0.15	0.15	0.10	0.10	0.05	0.05
M	0.30	0.25	0.20	0.10	0.10	0.05	0.05
N	0.30	0.25	0.25	0.15	0.15	0.05	0.05
O	0.30	0.25	0.25	0.15	0.15	0.05	0.05
X	0.15	0.15	0.15	0.10	0.10	0.05	0.05

注) 産業分類 "P","Q" の事業所は件数ゼロなので表示を省略。

表 6: 産業大分類コードと大分類産業部門名

産業分類	大分類の産業部門名
A	"Agriculture, hunting and forestry"
B	Fishing
C	Mining and quarrying
D	Manufacturing
E	"Electricity, gas and water supply"
F	Construction
G	"Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods"
H	Hotels and restaurants
I	"Transport, storage and communication"
J	Financial intermediation
K	"Real estate, renting and business activities"
L	Public administration and defence; compulsory social security
M	Education
N	Health and social work
O	"Other community, social and personal service activities"
P	Private households employing domestic staff and undifferentiated production activities of households for own use
Q	Extra-territorial organizations and bodies
X	Industry unknown

き事業所の属性に関する情報量が、第1番目の事業所ファイルを用いる実験 (この実験を以後は「実験1」とよぶ) と、第2番目の名寄せ済みファイルを用いる実験 (これを以後は「実験2」とよぶ) では異なり、一般に実験2のほうが前者に比べ情報量が多くなる。具体的には、前者の事業所ファイルを用いる実験では、事業所の産業分類や事業所の位置座標といった特定の事業所に固有の情報のみが得られるに過ぎないが、名寄せ済みファイルを用いる実験2では、事業所に固有の情報に加えて、当該事業所が属する企業全体の売上高階層などの情報も利用可能となる。これら二つの実験の目的は、事業所に関する属性の情報の追加による補完の精度の向上を確認することであり、事業所ファイルによる実験1におけるよりも、より豊富な情報を利用可能な名寄せ済みファイルを用いる実験2においていっそう高い補完の精度が期待される。

補完の精度は、本稿のこの節 (第4.1節) に以下に述べる実験手続きの詳細の第8項から第11項に示されるように、補完によって得られた従業者数の値、および従業者数の真の値である当該レコードに記録された従業者数の誤差および誤差率によっている。

以上に述べた補完の実験の手続きの概略は次の通りである。実験1、実験2のいずれにおいても、まず事業所レコードの一部に従業者数に擬似的な欠損値を生じさせ、従業者数に欠損値が生じていない一定の他の事業所レコードの集合からNIMにより選び取られた事業所レコードを「ドナー」とし、当該ドナーの従業者数により欠損値を補完する。実験1と実験2の違いは、NIMを適用する「従業者数に欠損値が生じていない他の事業所レコードの集合」(この集合を以後は「ドナーの候補集合」とよぶ) の違いである。ただし、いずれの実験においても、ドナーの候補集合に含まれる事業所レコードは、欠損値が生じている事業所レコードと同一の産業大分類に属する事業所のそれ (の全部あるいは一部) に限られ、この集合から、表1中の「事業所の位置座標」を用いて直線距離で最も近い事業所を「ドナー」として選び取る。以下に実験の手続きをより詳細に述べる。

1. 事業所ファイル、あるいは名寄せ済みファイルにおける、事業所の「従業者数」に欠損値を生じさせる確率を定める。欠損確率の定め方は、大別して二通りとし、まず事業所の属性にかかわらず一様に15%(ケース1)、30%(ケース2)、45%(ケース3) とした場合と、つぎに表5に従って事業所の産業大分類別と従業者階層別に欠損確率が変化する場合 (ケース4) を設定する。
   ただし、(ケース4) における全産業および全従業者階層の平均的欠損比率は、(ケース2) の30% にほぼ近い水準の、おおよそ27% となるように設定してある。この点は表7および表8の左上にある産業計・従業者数規模計の「欠損比率平均」の欄を参照されたい。なお、前者は実験1の (ケース4) における欠損比率の実績、後者は実験2のそれを示す。
   以下ではこの (ケース1) から (ケース4) の種類の別を、添え字 k ($k = 1, \ldots, 4$) で示す[2] 。
2. 以下の3から9までの手続きを、1回の試行として1,000回繰り返す。各試行には i ($i = 1, \ldots, 1000$) の番号[3] を付与する。
3. あらかじめ与えた上述の欠損確率に従い、試行 i において事業所ファイル、あるいは名寄せ済みファイルにおける先頭レコードから末尾レコードまでの間で、事業所の「従業者数」に、擬似的に欠損値を生じさせる。この擬似的な欠損値が生じた事業所 (以

[2] ケースの種類の別を示す k の添え字は、Kind of a case の先頭の文字より採用した。
[3] 試行の番号 i は Iteration の先頭の文字より採用した。

後は便宜的に「欠損事業所」とよぶ[4]) に先頭から番号 j ($j = 1, \ldots, J_{ik}$) を付与する。ただし、J_{ik} の値は、試行 i ごとにも異なる場合があるだけでなく、一般に上記 1 の (ケース 1) から (ケース 4) によっても変化する。

4. 事業所の従業者数が欠損となった欠損事業所 j に対応し、ドナーの候補集合として、従業者数が欠損している事業所と類似の属性を持ち、かつ欠損値が発生していない事業所 (レコード) の範囲を特定する。ただし、本稿における実験では、ドナーの候補集合は、事業所ファイル、名寄せ済みファイルのいずれを用いるかにより、次に述べる通りに二通りのものを設定した。

   事業所ファイルを用いた実験 1 における「ドナーの候補集合 $D_{jik}^{(1)}$」：試行 i のケース k において、従業者数について欠損値が発生している第 j 番目の欠損事業所と同一の産業大分類に属する事業所のうち、従業者数について欠損を生じていないもの。以下では実験 1 におけるドナーの候補集合を「$D_{jik}^{(1)}$」と記す。

   名寄せ済みファイルを用いた実験 2 における「ドナーの候補集合 $D_{jik}^{(2)}$」:試行 i のケース k において、従業者数について欠損値が発生している第 j 番目の欠損事業所と同一の産業大分類に属する事業所で、かつ当該事業所の企業の売上高階層と同一の売上高階層にある企業に名寄せされた事業所のうち、従業者数について欠損を生じていないもの。以下では実験 2 におけるドナーの候補集合を「$D_{jik}^{(2)}$」と記す。

5. ドナーの候補集合から、擬似的に欠損値を生じた事業所の位置座標から直線距離にして最も近い事業所のレコードをドナーとして選び出し、その事業所の従業員数によって、欠損値を補完する。併せてドナーとして選び出された事業所について、ドナーとなった回数を記録する。
6. 上の手続きによってすでにドナーとなった回数が 5 回に達している事業所があれば、その事業所はドナーの候補集合から除外し、上記 5 の手続きによりドナーを選び出す。
7. ドナーとして選ばれた事業所の従業者数を、従業者数の欠損値の補完に用いる。
8. 上の 4 において従業者数が擬似的に欠損値とされた事業所で報告されている本来の従業者数と、上の 7 で得られた補完値との誤差および誤差率を計算する。
9. 事業所ファイル、あるいは名寄せ済みファイルにおいて、擬似的に従業者数が欠損値とされたすべての事業所について上の 4 から 8 の手続きを終えたら、これを 1 回の試行として、この 1 回の試行において欠損値が生じたすべての事業所について計算した補完値の誤差および誤差率の平均値や標準偏差などの基本的統計量を記録する。
10. 上の 3 に戻り、新たに事業所ファイル、あるいは名寄せ済みファイルにおける先頭レコードから末尾レコードまでの間で、事業所の「従業者数」に、新たに擬似的に欠損値を生じさせ、以上の試行を 1,000 回繰り返す。ただし、毎回の試行 i ごとに、ドナーの候補集合 $D_{jik}^{(1)}$ および $D_{jik}^{(2)}$ に含まれるすべての事業所について、ドナーとなった回数をすべてゼロに戻してから毎回の試行 i を開始する。
11. 1,000 回の試行をすべて終えたら、上の 9 で計算された各試行における補完の誤差および誤差率の平均値や標準偏差などの基本的統計量について、1,000 回の試行全体にわたる平均値を計算し、実験 1 および実験 2 の各々におけるケース 1 からケース 4 について補完の精度を相互に比較する。

---

[4] 「欠損事業所」とは、当該事業所のレコードは事業所ファイル中に存在するが、当該事業所のレコードにおける従業者数が欠損値となっている場合を示すのであって、当該事業所の捕捉それ自体には問題がない点に注意されたい。

以上を要約すると、2種の実験はつぎのように示されよう。

実験 1 ：ドナーを事業所ファイルに含まれる、同一産業のレコードから構成される「ドナーの候補集合 $D_{jik}^{(1)}$ 」(試行 i のケース k において、j 番目の欠損事業所と同一の産業大分類に属する事業所のうち、従業者数について欠損を生じていないもの) から採用して補完する 1,000 回の試行

実験 2 ：ドナーを名寄せ済みファイルに含まれる、同一産業かつ同一売上高階層のレコードから構成される「ドナーの候補集合 $D_{jik}^{(2)}$ 」(試行 i のケース k において、第 j 番目の欠損事業所と同一の産業大分類に属する事業所で、かつ当該事業所の企業の売上高階層と同一の売上高階層にある企業に名寄せされた事業所のうち、従業者数について欠損を生じていないもの) から採用して補完する 1,000 回の試行

なお、次節 5 では、これら実験 1、実験 2 の「拡張」として、上記の、事業所の産業大分類別と従業者階層別に欠損確率が変化する (ケース 4) の場合についてのみ、別途実験を行った結果も併せて報告する。実験 1、実験 2 の拡張では、(ケース 4) の各々のドナーの候補集合 $D_{ji4}^{(1)}$, $D_{ji4}^{(2)}$ の部分集合[5] として、ドナーの候補集合に含まれる事業所について、試行 i における欠損事業所 j の従業者数階層と同一の従業者数階層に属するものに限定したドナーの候補集合 $D_{ji4}^{(1E)}$, $D_{ji4}^{(2E)}$ よりドナーを選ぶ実験を行った。これらの「拡張」として行った実験を以後は、各々「実験 1E」、「実験 2E」[6] とよぶ。いま、試行 i における欠損事業所 j について、この事業所の従業者数規模と同一の事業所の集合を、実験 1E で用いる事業所ファイルにおいては $C_{ji4}^{(1)}$ 、実験 2E で用いる名寄せ済みファイルにおいては $C_{ji4}^{(2)}$ としよう[7] 。これらの集合を用いれば、実験 1E、実験 2E は、つぎのように示されよう。

実験 1E ：ドナーの候補集合を $D_{ji4}^{(1E)} \equiv D_{ji4}^{(1)} \cap C_{ji4}^{(1)}$ として補完する 1,000 回の試行

実験 2E ：ドナーの候補集合を $D_{ji4}^{(2E)} \equiv D_{ji4}^{(2)} \cap C_{ji4}^{(2)}$ として補完する 1,000 回の試行

5 実験結果

この節では、実験結果を大別して次の二つの視点から比較検討する。

まず前節で述べた第 i 試行における、欠損値を生じている事業所 j の補完のためのドナーの候補集合を、$D_{jik}^{(1)}$ とした場合 (実験 1) と、$D_{jik}^{(2)}$ とした場合 (実験 2) の各々の実験結果を、欠損確率が産業分類・従業者数規模にかかわらず一様であるが、欠損確率の水準が 15%, 30%, 45% と変化する場合である (ケース 1) から (ケース 3) の場合、さらに欠損確率が産業大分類別および従業者数規模別に異なる (ケース 4) の場合について示す。すなわち、ドナーの候補集合の設定の方法は (ケース 1) から (ケース 4) 共通であるが、欠損確率の違いが補完の精度に与える影響をこれらの結果の比較により示すことがこれら実験の趣旨である。

つぎに、(ケース 4) の場合に限定して、これら実験の拡張として行った「実験 1E」と「実験 2E」の結果を、「実験 1」と「実験 2」の (ケース 4) における結果と比較する。この比較により、従業者数規模の情報の有無が補完の精度に与える影響を示すことができよう。

[5] 右下添え字の 4 は、欠損確率が (ケース 4) の場合に限定されている、すなわち $k = 4$ であることを示す。

[6] これらの実験の名前に付した "E" の文字は、従業者数階層である Employee Class, あるいは実験の「拡張」を意味する Extention の最初の文字より採用した。

[7] この集合の文字 C は、Class of Employee の最初の文字より採用した。

なお、「実験 1E」および「実験 2E」の第 i 番目の各々の試行において、擬似的に欠損値が発生している事業所の集合 (以後はこの集合を便宜的に「欠損事業所集合」とよぶ) を各々$M_{i4}^{(1E)}$, $M_{i4}^{(2E)}$ とし[8]、さらに (ケース 4) の下で「実験 1」および「実験 2」の第 i 番目の各々の試行における「欠損事業所集合」を各々$M_{i4}^{(1)}$, $M_{i4}^{(2)}$ とする。このとき

$$M_{i4}^{(1)} = M_{i4}^{(1E)}, \qquad (i = 1, \ldots, 1000)$$
$$M_{i4}^{(2)} = M_{i4}^{(2E)}, \qquad (i = 1, \ldots, 1000)$$

となるように実験が統御されているので、実験 1 と実験 1E の各々の補完の精度の間の差異は、純粋に「ドナーの候補集合」$D_{ji4}^{(1)}$, $D_{ji4}^{(1E)}$ (一般に $D_{ji4}^{(1)} \supseteq D_{ji4}^{(1E)}$ であるが、ほぼ 1 の確率で $D_{ji4}^{(1)} \supset D_{ji4}^{(1E)}$ が成り立つ) の差異に帰することができ、同様に実験 2 と実験 2E の各々の補完の精度の間の差異はいずれも、純粋に「ドナーの候補集合」$D_{ji4}^{(2)}$, $D_{ji4}^{(2E)}$ (一般に $D_{ji4}^{(2)} \supseteq D_{ji4}^{(2E)}$ で、ほぼ 1 の確率で $D_{ji4}^{(2)} \supset D_{ji4}^{(2E)}$ である) の差異に帰することができる。

さらに、ほぼ確率 1 で

$$M_{i4}^{(1)} \neq M_{\ell 4}^{(1E)}, \qquad (i \neq \ell,\ i, \ell = 1, \ldots, 1000)$$
$$M_{i4}^{(2)} \neq M_{\ell 4}^{(2E)}, \qquad (i \neq \ell,\ i, \ell = 1, \ldots, 1000)$$

が成立する。

なお、(ケース 4) の下での実験 1、実験 1E の欠損比率の実績は、表 7 に、同じく (ケース 4) の下での実験 2、実験 2E の欠損比率の実績は、表 8 に示した。

5.1 実験 1(E) と実験 2(E) の結果

実験 1 と実験 2、および実験 1E と実験 2E の結果の概略を表 9 に示した。ここでは主にこの表から読み取れる実験結果について考察する。

実験 1 および実験 2 のケース 1 からケース 4 の比較 :注目すべき結果は次の二点であろう。

まず第一点目として、欠損確率が事業所の産業分類や従業者数規模にかかわらず一様であれば、ケース 1 からケース 3 にかけての欠損確率の水準の変化は、誤差および誤差率に対して大きな影響を与えない、という点である。ケース 1 からケース 3 は欠損確率が事業所の産業大分類や従業者数規模にかかわらず一定であるが、これらケース 1 からケース 3 にかけては欠損確率が 3 倍に増加している。このようなケース 1 からケース 3 の結果を比較すると、興味深いことに、各試行における誤差あるいは誤差率の平均値ならびに標準偏差の 1,000 回の試行全体にわたる平均および標準偏差は大きく変化していないことがわかる。

第二点目として、欠損確率が事業所の産業分類や従業者数規模に応じて変化するケース 4 の場合、表 7 や表 8 に示すように、産業計および従業者規模計の平均的な欠損率が 0.268 と、ケース 2 のそれよりもやや小さくても、誤差の絶対値が、欠損率が一様である場合に比べて有意に大きくなる、という点である。ケース 1 からケース 3 の

[8] 「欠損事業所集合」を示すこれらの右下添え字の 1 番目は、試行 i における欠損事業所集合であることを示す。同様に右下添え字の 2 番目は、欠損値の発生確率が (ケース 4) である条件の下における欠損事業所集合であることを示す。この 2 番目の添え字が 4 であるのは、実験 1E および実験 2E が (ケース 4) の場合においてのみ行われているからである。

表 7: 事業所ファイルを用いる「実験 1 の (ケース 4)」における産業大分類別・事業所従業員規模別の欠損比率の実績

産業分類	統計量	事業所従業員規模							
		規模計	0〜4	5〜9	10〜19	20〜49	50〜99	100〜199	200〜
産業計	標本サイズ	208259	172615	18196	9323	5359	1598	728	440
	欠損比率平均	0.268	0.284	0.234	0.193	0.103	0.103	0.050	0.050
A	標本サイズ	5858	4999	534	241	59	20	4	1
	欠損比率平均	0.292	0.300	0.250	0.251	0.150	0.150	0.046	0.047
	同標準偏差	—	0.006	0.019	0.028	0.048	0.080	0.107	0.212
B	標本サイズ	405	387	16	1	1	0	0	0
	欠損比率平均	0.298	0.300	0.248	0.258	0.170	N/A	N/A	N/A
	同標準偏差	—	0.023	0.106	0.438	0.376	N/A	N/A	N/A
C	標本サイズ	809	688	71	24	19	3	3	1
	欠損比率平均	0.288	0.299	0.252	0.247	0.150	0.143	0.048	0.049
	同標準偏差	—	0.017	0.052	0.085	0.081	0.207	0.123	0.216
D	標本サイズ	21450	15292	2230	1600	1271	563	274	220
	欠損比率平均	0.143	0.150	0.150	0.150	0.100	0.100	0.050	0.050
	同標準偏差	—	0.003	0.007	0.009	0.008	0.012	0.013	0.014
E	標本サイズ	762	551	61	68	47	26	5	4
	欠損比率平均	0.144	0.150	0.150	0.151	0.099	0.102	0.047	0.049
	同標準偏差	—	0.015	0.045	0.043	0.042	0.059	0.095	0.109
F	標本サイズ	26213	21697	2385	1244	654	146	66	21
	欠損比率平均	0.284	0.300	0.250	0.200	0.100	0.100	0.049	0.049
	同標準偏差	—	0.003	0.009	0.012	0.012	0.025	0.026	0.046
G	標本サイズ	45907	36859	5185	2347	1142	259	84	31
	欠損比率平均	0.282	0.300	0.250	0.200	0.100	0.099	0.050	0.051
	同標準偏差	—	0.002	0.006	0.008	0.009	0.019	0.024	0.040
H	標本サイズ	9878	7667	1368	567	227	35	12	2
	欠損比率平均	0.282	0.300	0.250	0.200	0.100	0.103	0.048	0.044
	同標準偏差	—	0.005	0.012	0.017	0.020	0.052	0.064	0.149
I	標本サイズ	21061	17756	1658	834	504	151	87	71
	欠損比率平均	0.284	0.300	0.250	0.200	0.100	0.102	0.050	0.049
	同標準偏差	—	0.003	0.010	0.014	0.014	0.024	0.023	0.027
J	標本サイズ	4128	2792	617	401	204	60	37	17
	欠損比率平均	0.146	0.150	0.150	0.150	0.101	0.099	0.051	0.054
	同標準偏差	—	0.006	0.014	0.017	0.021	0.039	0.036	0.056
K	標本サイズ	41085	35450	2749	1481	958	263	126	58
	欠損比率平均	0.286	0.300	0.250	0.200	0.100	0.101	0.050	0.049
	同標準偏差	—	0.002	0.008	0.010	0.010	0.018	0.019	0.027
L	標本サイズ	36	19	4	3	3	0	1	6
	欠損比率平均	0.127	0.150	0.146	0.162	0.093	N/A	0.045	0.052
	同標準偏差	—	0.080	0.181	0.210	0.166	N/A	0.207	0.090
M	標本サイズ	1664	1448	103	39	44	25	5	0
	欠損比率平均	0.285	0.300	0.248	0.196	0.097	0.098	0.049	N/A
	同標準偏差	—	0.012	0.044	0.062	0.045	0.060	0.097	N/A
N	標本サイズ	13194	12117	714	246	92	19	5	1
	欠損比率平均	0.295	0.300	0.250	0.250	0.149	0.147	0.050	0.040
	同標準偏差	—	0.004	0.016	0.027	0.038	0.081	0.096	0.196
O	標本サイズ	15795	14879	501	227	134	28	19	7
	欠損比率平均	0.296	0.300	0.251	0.248	0.149	0.149	0.049	0.051
	同標準偏差	—	0.004	0.019	0.031	0.032	0.070	0.048	0.083
X	標本サイズ	14	14	0	0	0	0	0	0
	欠損比率平均	0.152	0.152	N/A	N/A	N/A	N/A	N/A	N/A
	同標準偏差	—	0.097	N/A	N/A	N/A	N/A	N/A	N/A

注 1) 産業分類 "P","Q" の事業所は件数ゼロなので表示を省略。
注 2) 産業計・規模計においては欠損比率の標準偏差の表示を省略。
注 3) 表中の "N/A" は標本サイズがゼロのため計算できないことを示す。
注 4)「実験 1」では事業所の緯度・経度が欠損しているレコードを除外してあるので、産業計・規模計の標本サイズ 208, 259 は表 3 のレコード数合計 255, 127 よりも除外した分だけ小さい。

表 8: 名寄せ済みファイルを用いる「実験 2 の (ケース 4)」における産業大分類別・事業所従業員規模別の欠損比率の実績

産業分類	統計量	事業所従業員規模							
		規模計	0〜4	5〜9	10〜19	20〜49	50〜99	100〜199	200〜
産業計	標本サイズ	207718	172195	18124	9291	5346	1595	728	439
	欠損比率平均	0.268	0.284	0.234	0.193	0.103	0.102	0.050	0.050
A	標本サイズ	5837	4981	533	239	59	20	4	1
	欠損比率平均	0.291	0.300	0.250	0.250	0.150	0.148	0.046	0.048
	同標準偏差	—	0.006	0.019	0.029	0.044	0.078	0.103	0.214
B	標本サイズ	405	387	16	1	1	0	0	0
	欠損比率平均	0.297	0.300	0.248	0.255	0.149	N/A	N/A	N/A
	同標準偏差	—	0.023	0.104	0.436	0.356	N/A	N/A	N/A
C	標本サイズ	809	688	71	24	19	3	3	1
	欠損比率平均	0.288	0.300	0.250	0.247	0.148	0.161	0.050	0.049
	同標準偏差	—	0.017	0.051	0.088	0.081	0.208	0.124	0.216
D	標本サイズ	21381	15251	2213	1592	1268	563	274	220
	欠損比率平均	0.143	0.150	0.150	0.150	0.100	0.100	0.050	0.050
	同標準偏差	—	0.003	0.008	0.009	0.008	0.013	0.013	0.015
E	標本サイズ	762	551	61	68	47	26	5	4
	欠損比率平均	0.144	0.150	0.149	0.150	0.102	0.100	0.049	0.046
	同標準偏差	—	0.015	0.046	0.043	0.044	0.059	0.098	0.105
F	標本サイズ	26207	21692	2384	1244	654	146	66	21
	欠損比率平均	0.284	0.300	0.250	0.200	0.100	0.101	0.050	0.049
	同標準偏差	—	0.003	0.009	0.011	0.012	0.026	0.027	0.047
G	標本サイズ	45767	36732	5182	2339	1140	259	84	31
	欠損比率平均	0.282	0.300	0.250	0.200	0.100	0.100	0.050	0.049
	同標準偏差	—	0.002	0.006	0.008	0.009	0.018	0.025	0.038
H	標本サイズ	9695	7538	1334	556	221	32	12	2
	欠損比率平均	0.282	0.300	0.251	0.200	0.100	0.102	0.053	0.045
	同標準偏差	—	0.005	0.013	0.017	0.021	0.054	0.065	0.149
I	標本サイズ	21057	17753	1658	833	504	151	87	71
	欠損比率平均	0.284	0.300	0.251	0.200	0.100	0.099	0.050	0.051
	同標準偏差	—	0.003	0.011	0.014	0.013	0.024	0.023	0.026
J	標本サイズ	4128	2792	617	401	204	60	37	17
	欠損比率平均	0.146	0.150	0.150	0.151	0.100	0.099	0.050	0.053
	同標準偏差	—	0.007	0.014	0.018	0.021	0.037	0.034	0.055
K	標本サイズ	41022	35393	2746	1480	956	263	126	58
	欠損比率平均	0.286	0.300	0.250	0.200	0.100	0.101	0.051	0.049
	同標準偏差	—	0.002	0.008	0.010	0.009	0.019	0.020	0.029
L	標本サイズ	36	19	4	3	3	0	1	6
	欠損比率平均	0.125	0.147	0.155	0.147	0.096	N/A	0.053	0.052
	同標準偏差	—	0.080	0.181	0.207	0.171	N/A	0.224	0.094
M	標本サイズ	1662	1446	103	39	44	25	5	0
	欠損比率平均	0.285	0.300	0.252	0.199	0.101	0.102	0.051	N/A
	同標準偏差	—	0.012	0.042	0.065	0.044	0.059	0.098	N/A
N	標本サイズ	13191	12115	714	245	92	19	5	1
	欠損比率平均	0.295	0.300	0.250	0.250	0.151	0.146	0.053	0.045
	同標準偏差	—	0.004	0.015	0.027	0.037	0.080	0.102	0.207
O	標本サイズ	15745	14843	488	227	134	28	19	6
	欠損比率平均	0.296	0.300	0.250	0.251	0.150	0.149	0.050	0.048
	同標準偏差	—	0.004	0.020	0.029	0.031	0.067	0.049	0.089
X	標本サイズ	14	14	0	0	0	0	0	0
	欠損比率平均	0.150	0.150	N/A	N/A	N/A	N/A	N/A	N/A
	同標準偏差	—	0.097	N/A	N/A	N/A	N/A	N/A	N/A

注 1) 産業分類 "P","Q" の事業所は件数ゼロなので表示を省略。
注 2) 産業計・規模計においては欠損比率の標準偏差の表示を省略。
注 3) 表中の "N/A" は標本サイズがゼロのため計算できないことを示す。
注 4)「実験 2」では事業所の緯度・経度が欠損しているレコードを除外してあるので、産業計・規模計の標本サイズ 207,718 は表 4 のレコード数合計 226,426 よりも除外した分だけ小さい。

表 9: 実験 1(E) と実験 2(E) の比較：地理的近接の同一産業大分類事業所による補完誤差 (率) の記述統計

ケース	各試行における誤差 (率) の統計量	誤差				誤差率			
		実験 1		実験 2		実験 1		実験 2	
		1,000 回の試行全体の		1,000 回の試行全体の		1,000 回の試行全体の		1,000 回の試行全体の	
		平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差
ケース 1 (15%)	標本サイズ	31238.8	17.5	31156.8	18.9	30554.3	29.8	30480.6	28.8
	平均値	0.262	0.804	−0.094	0.639	−9.609	1.176	−1.661	0.527
	標準偏差	134.797	7.408	106.793	8.488	202.925	86.043	74.315	67.382
ケース 2 (30%)	標本サイズ	62477.6	29.8	62312.9	31.3	61109.1	41.7	60961.4	40.7
	平均値	0.243	0.640	−0.192	0.485	−9.978	0.879	−1.757	0.434
	標準偏差	137.439	5.028	109.370	5.648	215.860	64.137	91.072	59.956
ケース 3 (45%)	標本サイズ	93716.3	37.7	93467.6	38.6	91664.3	47.5	91438.1	47.9
	平均値	0.161	0.569	−0.266	0.472	−10.566	0.761	−1.886	0.388
	標準偏差	140.143	4.192	111.994	4.674	224.234	50.904	106.284	52.380
ケース 4	標本サイズ	55811.2	75.1	55664.5	73.1	54643.7	78.8	54513.3	76.2
	平均値	−2.561	0.533	−1.092	0.345	−10.241	0.838	−1.414	0.296
	標準偏差	112.811	4.695	76.771	5.813	192.802	58.406	56.908	49.800
		実験 1E		実験 2E		実験 1E		実験 2E	
ケース 4	標本サイズ	55810.6	75.1	55648.1	73.1	54643.0	78.9	54496.9	76.2
	平均値	0.292	0.505	−0.133	0.335	−3.882	0.437	−0.679	0.118
	標準偏差	107.732	4.743	75.415	5.438	93.121	34.230	16.041	23.123

注 1) 表側の「ケース 1」から「ケース 4」は擬似的に発生させる欠損確率を示す。「ケース 1」から「ケース 3 では産業大分類や従業員数規模にかかわりなく一定の欠損確率とし、これを括弧内に示す。「ケース 4」では産業大分類や従業者数規模により、表 5 に示すとおりに欠損確率が変化するとした。

注 2) 表側の「標本サイズ」は各試行における欠損事業所数を示す。

注 3) 表側の「平均値」は各試行における補完の誤差 (率) の平均値を示す。たとえば実験 1 のケース 4 では、各試行における誤差の平均値を、1,000 回の試行全体を通して平均すると −2.561、その標準偏差は 0.533、同じく各試行における誤差率の平均値を、1,000 回の試行全体を通して平均すると −10.241、その標準偏差は 0.838 となることを示す。

注 4) 表側の「標準偏差」は各試行における補完の誤差 (率) の標準偏差を示す。たとえば実験 1 のケース 4 では、各試行における誤差の標準偏差を、1,000 回の試行全体を通して平均すると 112.811、その標準偏差は 4.695、同じく各試行における誤差率の標準偏差を、1,000 回の試行全体を通して平均すると 192.802、その標準偏差は 58.406 となることを示す。

場合、誤差の平均値の 1,000 回にわたる試行全体の平均値は、実験 1 では 0.161 から 0.262 の範囲、実験 2 では −0.266 から −0.094 の範囲であるのに対し、ケース 4 の場合、実験 1 では −2.561 、実験 2 では −1.0992 と、その絶対値は有意に大きくなっている。

一方で、誤差率は、ケース 1 からケース 3 の場合と、ケース 4 の場合を比較しても有意な差は見られない。

実験 1 と実験 2 の比較 ：注目すべきは、実験 2 において各試行における誤差率の平均値の 1,000 にわたる平均値と標準偏差について、平均値の絶対値は劇的に減少し、さらに標準偏差も減少している点である。

一方で、誤差については、こうした若干の傾向は見られるものの、その変化の傾向は誤差率のそれに比べ必ずしも明確ではない。

実験 1 と実験 1E、実験 2 と実験 2E の比較 ：誤差、および誤差率のいずれにおいても、各試行における平均値の、1,000 回の試行全体にわたる平均値の絶対値は有意に小さくなっており、補完の精度が明らかに向上していることが読み取れる。

一方で、各試行における平均値の、1,000 回の試行全体にわたる標準偏差には明確な傾向は現れていない。

誤差と誤差率の比較 ：注目すべきは次の二点であろう。

まず、誤差率は実験 1(E) および実験 2(E) のいずれのケースにおいても、有意に負値となっているが、誤差の絶対値が有意に正値となるのは実験 2 のケース 4 の場合に限られる、という点である。

つぎに、誤差あるいは誤差率の絶対値が有意に正値となる場合には、すべての場合において、誤差あるいは誤差率が有意に負値となっている点である。これは事業所の従業者数規模の分布が、多数の小規模事業所と、ごく少数の大規模事業所という、事業所規模の形状の特殊性によるものと考えられる。

実験結果のまとめ ：以上の結果をまとめると、注目すべきは以下の三点であろう。

第一点目は、実験 1 と実験 1E、あるいは実験 2 と実験 2E を各々比較することにより、従業員規模に関する情報の追加が補完の精度に大きく貢献する点が明らかとなる。これは、調査項目として従業者数における欠損値が多発する場合であっても、その階層に関する情報があれば、補完の精度を向上させることが可能であることを示唆している。この点は、誤差および誤差率のいずれについても当てはまる。

第二点目は、実験 1 と実験 2、あるいは実験 1E と実験 2E を各々比較することにより、事業所に関する情報だけでなく、当該事業所が属する企業の売上高階層などの情報の追加が、やはり補完の制度の向上に貢献することが明らかとなる。この点は、とくに誤差率について当てはまる。

第三点目は、ケース 1 からケース 3 までの結果とケース 4 の結果を比較することにより、欠損値の発生確率が事業所の産業や規模に依存して変化する場合には、それが一様である場合に比べ、補完の精度が悪くなる、という点が明らかとなる。この点は、とくに誤差に当てはまる。欠損値の発生状況が、調査対象の個体属性により変化することは一般的に広く経験されるのは周知の通りである。従って、「ドナーの候補集合」

を限られた情報により設定し補完を行った場合に十分な精度が確保できない恐れがある。こうした場合、上記の第一点目や第二点目の結果は、ある特定の調査項目の数値に欠損が生じている場合であっても、当該項目に関する階層の情報を追加したり、個体が属するグループに関する追加的情報により「ドナーの候補集合」を設定することが、補完の精度を向上させるよい決め手となることを示していると言えよう。

なお、参考までに表10、表12には各々、実験1E、実験2Eにおける補完の精度に関する結果を産業大分類別に示した。さらに、表11、表13には各々、実験1E、実験2Eにおける補完のドナーとして選ばれた事業所までの地理的距離の状況を産業大分類別に示してある。

表10、表12の補完の精度には、産業大分類の違いによる大きな差異は見られないが、表11、表13の補完のドナーとして選ばれた事業所までの地理的距離の状況には、産業大分類の違いによる差異が明らかに見られる。しかし、この違いは、クラスターを形成するといった産業立地の特性だけでなく、標本サイズの違いにも依存しているようにも見えることから、さらなる分析が必要である。

6　おわりに

本稿ではフィンランドの2002年ビジネスレジスターのデータセットを用い、NIMによる補完の実験を行った。NIMによる補完の基本は、基本的にある特定の項目について欠損値が生じている個体と共通の属性を持つ別の個体集合「ドナーの候補集合」を得て、その集合の中から、欠損値が生じている当該個体と地理的に最近接している個体を選び、その最近接している個体の値を用いるという点である。本稿では、この「ドナーの候補集合」の設定において、補完の精度が追加的な情報により向上することを示すことができた。ここで言う、追加的な情報とは大別して次の二点である。

第一点目は、欠損値が生じている項目の数値情報に替えて、当該数値が属する階層に関する情報である。本稿で確認できたように、従業者数に欠損値があっても、これに替えて従業者数階層の情報が得られるのであれば、補完の精度が飛躍的に向上することが明らかにされた。この点は、調査実施についても次に述べる重要な示唆を与えている。すなわち、数値を記入してもらう調査項目だけでなく、当該数値が属する階層を選択して答えてもらう形式の調査項目を併用することにより、前者のについて欠損が生じても、後者の情報が得られない場合に比べ、より高い精度の補完が可能となる点である。

第二点目、欠損が生じた個体が属するグループに関する情報である。本稿では、事業所の従業者数について欠損が生じた場合、当該の欠損事業所の産業分類に加え、当該の欠損事業所が属する企業の売上高階層の情報を付加して「ドナーの候補集合」を設定することにより、この売上高階層の情報を用いない場合に比べ、補完の精度がより向上することが確認できた。

本稿では、ビジネスレジスターの事業所に関する調査項目である従業者数についての欠損値の補完の精度を検証してきたが、同様の議論は、事業所に関する調査項目に限らず、たとえばある企業に雇用される従業者に関する属性の欠損値の補完についても同様に当てはまるであろう。

以上の点を実際に施すには、調査全体の設計を考慮することが必要である。経験が示すところによれば、欠損値が生じ易い数値項目にはある一定の傾向があることから、そうし

た傾向を持つ数値項目を調査票に盛り込む場合、当該数値項目の補完の精度を確保するに足りるであろう、追加的な情報も得られるような調査の設計が要請されることを、本稿の実験結果は示していると言えよう。

表 10: 実験 1E：地理的近接の同一産業大分類事業所による補完誤差 (率) の記述統計

同一産業に属す、地理的距離が近い一事業所による補完					
		誤差		誤差率	
産業	統計量	平均	標準偏差	平均	標準偏差
Total	標本サイズ	55810.6	75.1	54643.0	78.9
	平均値	0.292	0.505	−3.882	0.437
	標準偏差	107.732	4.743	93.121	34.230
A	標本サイズ	1707.7	34.5	1688.7	34.3
	平均値	−0.360	2.052	−2.170	1.189
	標準偏差	72.355	24.883	24.362	27.419
B	標本サイズ	120.1	9.3	118.6	9.2
	平均値	−0.038	0.195	−1.393	0.497
	標準偏差	1.809	0.274	4.504	2.168
C	標本サイズ	233.1	12.8	228.7	12.6
	平均値	−0.181	0.843	−2.246	1.352
	標準偏差	9.615	6.928	11.135	16.943
D	標本サイズ	3075.5	48.4	3050.8	48.3
	平均値	0.312	2.050	−1.811	0.592
	標準偏差	106.033	30.533	19.573	23.038
E	標本サイズ	109.9	9.8	90.7	9.1
	平均値	1.543	17.025	−5.807	4.006
	標準偏差	138.960	106.767	30.651	23.176
F	標本サイズ	7436.2	69.0	7423.8	69.0
	平均値	−0.079	0.269	−1.305	0.088
	標準偏差	21.258	9.935	5.698	3.037
G	標本サイズ	12967.7	83.2	12570.1	82.0
	平均値	0.195	0.936	−5.568	1.069
	標準偏差	96.908	8.955	98.364	60.415
H	標本サイズ	2784.9	43.7	2769.2	43.6
	平均値	2.661	5.155	−18.225	4.326
	標準偏差	245.410	10.405	209.978	78.367
I	標本サイズ	5982.6	60.1	5952.1	60.1
	平均値	0.113	2.358	−5.065	2.037
	標準偏差	166.543	14.305	131.036	88.416
J	標本サイズ	600.8	21.5	445.1	19.2
	平均値	4.695	17.509	−11.492	15.616
	標準偏差	404.535	61.937	144.422	300.804
K	標本サイズ	11747.4	80.7	11314.3	80.2
	平均値	0.140	0.402	−1.710	0.279
	標準偏差	40.206	7.849	18.659	21.828
L	標本サイズ	4.5	1.9	3.6	1.7
	平均値	−0.782	43.947	−1.417	3.883
	標準偏差	37.969	73.209	3.076	5.607
M	標本サイズ	474.8	18.3	463.8	18.1
	平均値	0.272	4.072	−2.513	1.323
	標準偏差	71.753	42.287	16.606	17.669
N	標本サイズ	3890.1	51.0	3880.5	50.9
	平均値	0.052	0.342	−1.939	0.374
	標準偏差	19.823	1.966	21.572	9.651
O	標本サイズ	4673.1	57.7	4641.0	57.6
	平均値	0.209	0.761	−1.778	0.813
	標準偏差	46.945	10.094	42.553	37.479
X	標本サイズ	2.1	1.4	2.1	1.4
	平均値	0.028	0.328	−0.308	0.845
	標準偏差	0.268	0.262	0.626	0.843

注) 産業分類 "P","Q" の事業所は件数ゼロなので表示を省略。

表 11: 実験 1E：地理的近接の同一産業大分類ドナー事業所までの「距離」の記述統計

同一産業に属す、地理的距離が近い一事業所による補完					
産業分類	統計量	平均	標準偏差	最小値	最大値
Total	標本サイズ	55811.2	75.1	55591	56020
	平均値	973.63	20.34	920.13	1036.58
	標準偏差	4403.54	271.16	3610.10	5280.83
A	標本サイズ	1707.8	34.5	1610	1827
	平均値	2551.27	85.29	2331.45	2888.98
	標準偏差	3399.10	314.93	2807.97	4609.51
B	標本サイズ	120.5	9.3	95	151
	平均値	10374.14	1313.80	6789.26	15080.96
	標準偏差	14229.35	3023.86	7900.53	26772.37
C	標本サイズ	233.2	12.8	190	282
	平均値	7563.69	625.60	5917.85	10004.53
	標準偏差	9328.91	2199.36	5934.01	17998.18
D	標本サイズ	3075.5	48.4	2929	3239
	平均値	956.11	62.75	752.60	1199.25
	標準偏差	3516.72	774.79	2089.76	6307.33
E	標本サイズ	109.9	9.8	79	140
	平均値	5522.31	891.87	2831.94	9046.00
	標準偏差	8509.41	2144.77	4395.75	21224.56
F	標本サイズ	7436.2	69.0	7126	7651
	平均値	1029.81	46.98	892.64	1171.90
	標準偏差	3818.06	651.19	2195.04	5667.21
G	標本サイズ	12967.7	83.2	12684	13202
	平均値	677.17	38.50	560.35	804.46
	標準偏差	4073.71	574.18	2856.80	6643.43
H	標本サイズ	2784.9	43.7	2658	2920
	平均値	1238.15	97.05	988.08	1632.27
	標準偏差	4895.15	1121.71	2725.13	9444.45
I	標本サイズ	5982.6	60.1	5780	6238
	平均値	1063.28	50.18	917.40	1285.43
	標準偏差	3625.43	654.74	2340.99	5919.61
J	標本サイズ	600.8	21.5	532	668
	平均値	1426.17	200.96	911.50	2272.39
	標準偏差	4711.75	1271.21	2455.47	11223.86
K	標本サイズ	11747.4	80.7	11488	12032
	平均値	643.27	45.14	534.83	822.12
	標準偏差	4619.63	650.11	2814.78	7526.84
L	標本サイズ	4.6	1.9	0	12
	平均値	19955.39	34315.07	0.00	304270.35
	標準偏差	35897.66	53731.06	0.00	200891.67
M	標本サイズ	474.8	18.3	421	539
	平均値	2874.41	386.87	1849.94	4367.43
	標準偏差	7958.97	1605.37	4353.59	14278.70
N	標本サイズ	3890.1	51.0	3735	4064
	平均値	848.80	61.22	688.01	1126.97
	標準偏差	3915.74	884.25	2430.18	7137.77
O	標本サイズ	4673.1	57.7	4525	4866
	平均値	824.80	50.96	684.68	1033.00
	標準偏差	3403.34	695.04	2199.57	6355.84
X	標本サイズ	2.1	1.4	0	8
	平均値	54465.19	37191.18	0.00	226313.11
	標準偏差	25432.91	26395.76	0.00	98930.23

注) 産業分類 "P","Q" の事業所は件数ゼロなので表示を省略。

表 12: 実験 2E：地理的近接の同一産業大分類事業所による補完誤差 (率) の記述統計

同一産業・企業の同一売上階層に属す、地理的距離が近い一事業所による補完					
		誤差		誤差率	
産業	統計量	平均	標準偏差	平均	標準偏差
Total	標本サイズ	55648.1	73.1	54496.9	76.2
	平均値	−0.133	0.335	−0.679	0.118
	標準偏差	75.415	5.438	16.041	23.123
A	標本サイズ	1697.8	34.4	1680.3	34.2
	平均値	0.154	1.682	−0.725	0.319
	標準偏差	62.265	23.627	10.445	8.486
B	標本サイズ	119.7	9.0	118.2	9.0
	平均値	−0.034	0.129	−0.341	0.123
	標準偏差	1.193	0.274	1.257	0.215
C	標本サイズ	231.5	13.0	227.0	12.9
	平均値	−0.241	0.640	−0.552	0.244
	標準偏差	8.047	2.944	2.684	1.748
D	標本サイズ	3065.6	49.9	3041.5	49.6
	平均値	−0.475	1.996	−0.640	0.355
	標準偏差	103.469	30.254	11.377	17.251
E	標本サイズ	108.5	9.4	89.5	8.7
	平均値	−1.363	17.263	−2.632	2.790
	標準偏差	122.340	108.356	17.791	21.159
F	標本サイズ	7437.2	68.8	7425.3	68.7
	平均値	−0.129	0.232	−0.332	0.038
	標準偏差	17.290	8.293	1.951	1.688
G	標本サイズ	12924.9	85.3	12531.3	84.1
	平均値	−0.236	0.739	−0.745	0.039
	標準偏差	77.655	9.851	4.258	0.773
H	標本サイズ	2732.2	45.3	2720.3	45.3
	平均値	−1.071	1.224	−0.594	0.099
	標準偏差	57.688	8.521	4.944	1.855
I	標本サイズ	5980.9	60.5	5950.9	60.3
	平均値	0.323	1.408	−0.551	0.117
	標準偏差	101.283	16.777	7.929	4.615
J	標本サイズ	601.1	22.8	445.4	19.7
	平均値	5.151	17.373	−12.233	14.277
	標準偏差	405.781	60.447	147.326	263.532
K	標本サイズ	11728.9	81.5	11299.6	79.9
	平均値	−0.165	0.309	−0.618	0.036
	標準偏差	28.233	7.640	3.523	1.327
L	標本サイズ	4.4	2.0	3.5	1.8
	平均値	3.434	45.341	−1.549	4.124
	標準偏差	38.056	73.810	3.264	5.715
M	標本サイズ	472.1	18.7	461.2	18.4
	平均値	−0.651	1.320	−0.767	0.487
	標準偏差	24.182	9.309	7.016	6.667
N	標本サイズ	3886.4	49.1	3876.8	49.0
	平均値	−0.154	0.246	−0.478	0.053
	標準偏差	13.109	2.413	3.215	0.724
O	標本サイズ	4655.4	57.9	4624.6	57.6
	平均値	−0.272	0.567	−0.456	0.044
	標準偏差	32.781	10.970	2.712	0.838
X	標本サイズ	1.6	1.2	1.6	1.2
	平均値	0.005	0.292	−0.153	0.552
	標準偏差	0.174	0.216	0.285	0.458

注) 産業分類 "P","Q" の事業所は件数ゼロなので表示を省略。

表 13: 実験 2E：地理的近接の同一産業大分類ドナー事業所までの「距離」の記述統計

同一産業・企業の同一売上階層に属す、地理的距離が近い一事業所による補完					
産業分類	統計量	平均	標準偏差	最小値	最大値
Total	標本サイズ	55664.5	73.1	55441	55933
	平均値	2244.36	32.54	2136.26	2359.86
	標準偏差	7576.30	305.87	6637.34	9272.91
A	標本サイズ	1700.4	34.4	1592	1835
	平均値	7662.89	275.22	6896.61	8695.54
	標準偏差	10433.10	1263.24	7606.35	18527.13
B	標本サイズ	120.4	9.1	93	153
	平均値	26802.34	3971.41	18949.24	46349.35
	標準偏差	36197.59	9740.40	19770.04	90361.82
C	標本サイズ	233.3	13.1	189	276
	平均値	22544.89	1876.01	17597.31	28894.54
	標準偏差	27934.46	5375.73	14862.16	47991.77
D	標本サイズ	3066.2	49.9	2908	3216
	平均値	2139.26	100.89	1823.15	2609.10
	標準偏差	5576.98	928.01	3808.25	9308.46
E	標本サイズ	109.6	9.5	76	141
	平均値	23199.82	2984.22	14767.07	36816.41
	標準偏差	29309.67	6886.43	16073.20	74337.06
F	標本サイズ	7438.2	68.8	7221	7670
	平均値	2106.71	64.71	1910.01	2314.00
	標準偏差	5413.31	474.27	4148.39	7276.66
G	標本サイズ	12926.0	85.2	12676	13204
	平均値	1303.92	42.10	1167.87	1422.28
	標準偏差	4790.57	383.26	3799.41	6065.33
H	標本サイズ	2733.1	45.3	2560	2894
	平均値	3612.06	186.50	3109.38	4331.58
	標準偏差	9377.53	876.57	7320.66	15584.92
I	標本サイズ	5982.1	60.5	5773	6149
	平均値	2461.26	88.39	2231.38	2853.04
	標準偏差	6559.49	808.76	4806.25	12888.81
J	標本サイズ	601.1	22.8	537	673
	平均値	1805.89	430.84	935.09	3771.55
	標準偏差	9317.78	4824.20	2757.50	29044.39
K	標本サイズ	11729.2	81.5	11454	11994
	平均値	1131.99	48.21	1009.37	1378.51
	標準偏差	5265.21	523.66	3958.80	6901.34
L	標本サイズ	4.5	2.0	0	11
	平均値	20610.31	36627.43	0.00	304270.35
	標準偏差	36166.44	54452.67	0.00	200925.23
M	標本サイズ	474.3	18.8	424	528
	平均値	9798.29	940.08	6467.50	13034.18
	標準偏差	21020.71	2837.35	13480.94	33265.90
N	標本サイズ	3887.6	49.1	3729	4065
	平均値	2178.33	123.12	1811.39	2675.58
	標準偏差	7221.01	722.83	5503.51	9889.72
O	標本サイズ	4657.1	58.0	4474	4835
	平均値	2031.74	108.26	1758.70	2424.38
	標準偏差	7380.71	712.05	5653.65	12009.95
X	標本サイズ	1.6	1.2	0	6
	平均値	67310.09	47207.93	0.00	270019.21
	標準偏差	20160.33	26306.51	0.00	145650.04

注) 産業分類 "P","Q" の事業所は件数ゼロなので表示を省略。

参考文献

[1] Andridge, R. and R. Little (2010); "A Review of Hot Deck Imputaion for Survey Non-response," *International Statistical Review*, **78**(1), pp.40-64.

[2] Bankier, M., A.-M. Houle and M. Luc (1997); "1996 Canadian Census Demographic Variables Imputation," *Proceedings of the Survey Methods Section, SSC Annual Meeting, June 1997.*

[3] Bankier, M., P. Poirier and M. Lachance (2001); "Efficient Methodology within the Candadian Census Edit and Imputation System (CANCEIS)," *Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001.*

[4] Chen, J. and J. Shao (2000); "Nearest Neighbor Imputation for Survey Data," *Journal of Official Statistics*, **16**(2), pp.113-131.

[5] Fellegi, I. and D. Holt (1976); "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, **71**(353), pp.17-35.

[6] Little, R. and D. Rubin (1987); *Statistical Analysis with Missing Data*, John Wiley and Sons, Inc.

[7] Little, R. and D. Rubin (2002); *Statistical Analysis with Missing Data*, Second Edition, New Jersey: John Wiley and Sons, Inc.

[8] Rubin, D. (1976); "Inference and Missing Data," *Biometrika*, **63**(3), pp.581-592.

[9] Rubin, D. (1987); *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons, Inc.

[10] Rubin, D. (1996); "Multiple Imputation after 18+ Years", *Journal of the Americal Statistical Association*, **91**(434), pp.473-489.

[11] Statistics Canada (prepared by M. Bankier), (1999); "Experience with the New Imputation Methodology Used in the 1996 Canadadian Census with Extensions for Future Censuses," *Statistical Commission and Economic Commission for Europe Working Paper*, No.24.

[12] Statistics Canada (prepared by M. Bankier, P. Mason and P. Poirier), (2002); "Imputation of Demographic Variables from the 2001 Canadian Census of Population," *Statistical Commission and Economic Commission for Europe Working Paper*, No.25.

[13] Winkler, W. and B. Chen (2001); "Extending the Fellegi-Holt Model of Statistical Data Editing," *Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001.*

[14] 高橋将宣・伊藤孝之 (2013); 「経済調査における売上高の欠測値補定方法について」『統計研究彙報』第 70 号, pp.19-86.

[15] 森博美・菅幹雄 (2012);「事業所母集団データベースの更新情報等を活用したレジスター統計に関する研究」『平成 23 年度統計研修所共同研究報告書』総務省統計局統計研修所