

平成31年個人企業経済調査 ～経理項目の補完に向けて～

**(独) 統計センター
統計技術研究課**

NSTAC

目 次

1. 研究目的
2. 現在の検討状況と今後の予定
3. 補完対象項目
4. 研究の基本方針
5. 補完にあたり考慮すべき事項
6. 比率ホットデスク
7. 補完データの時点調整
8. 補完処理の流れ
9. 経理項目
10. 従業者数関連項目
11. 設備投資関連項目
12. 専従者給与について
13. 今後の課題

1. 研究目的

- 売上金額、仕入金額、経費等の主要な**量的変数（金額）**で、GDP推計にも使用される調査項目について欠測を補完する
- 調査票未回収分（**単位欠測**）及び一部非回答分（**項目欠測**）の**両方を対象**とする
- H28経済センサス-活動調査と同様の補完方法（ロバスト比率補完）を適用手法の候補とする

要件

- 必要項目は全て、何らかの方法で欠測を補完する
- 時点の違うデータを使用する場合は、そのまま横置きをせず、時点の違いを調整する

2. 現在の検討状況と今後の予定

- ◆ 本年4月に補完対象項目の提示を受け、平成24年経済センサス-活動調査データを用いて大まかな事前分析を行い、補完案の概要を固めたところ
- ◆ 7月から、平成28年経済センサス-活動調査データが利用可能になり、より詳細レベルで確認のための分析を行っている
- ◆ 9月までには、補完処理の流れを固め、その後補完クラスの確定を含めた各論部分の検討を行う

3. 補完対象項目

05	売上金額
06	費用総額*
07	期首棚卸高
08	仕入高
09	期末棚卸高
10	経費計
11	給料賃金
12	地代家賃
13	減価償却費
14	租税公課
15	損害保険料
16	福利厚生費
17	外注工賃
18	利子割引料

19	設備投資（新規設備取得）
20	うち車両 機械 工具 器具 備品
201	最も取得額の多い時期
21	設備投資（中古設備取得）
32	専従者給与
04	従業者数
041	事業主の家族で無給
042	常用雇用者
043	パート・アルバイト
044	臨時雇用者

- *印のある「06 費用総額」は補完対象ではないが、補完作業のために中間データとして必要となる
- 項目番号が紫のものは、H28経済センサス-活動調査に同じ項目が存在する
- 項目名が太字でないものは、H30までの個人企業経済調査では対象項目ではない

4. 研究の基本方針

1. 補完には、新調査データの他に、時点の異なる二種類のデータを使用する

H28経済センサス-活動調査 **【H28センサス】**

H29事業所母集団DBから作成する抽出用母集団名簿
(H28センサス以降の新設企業情報を含む) **【H29DB】**

H31個人企業経済調査 **【H31新調査】**

2. 新調査データに欠測があり、他の二種類のデータに同一企業の情報が含まれている場合は、まずそのデータを用いて補完する (人数データ以外は時点の違いを調整する)
3. 集計スケジュールを考慮し、補完クラス（補完作業を行うデータのグループ）の選択等、既存データで可能な作業は、新調査の調査票回収前に行う

補完クラスについて

- 無回答により起こる結果数値のバイアスを減らし、推定精度を上げるために設定する
- クラスは全ての調査単位について欠測のない変数により、クラス内の補完対象ができるだけ均質で、クラス間の差異は大きくなるように設定する
- クラスを構築する変数は、回答・無回答の傾向に関係するものを使用する

例) 大企業と中小企業で、回答する企業の割合が異なる場合
補完クラスを規模に関する変数、例えば従業員数等を用いて設定し、クラス内の回答割合をできるだけ均一にする

5. 補完にあたり考慮すべき事項

■ 事前分析データ

モデルを使用する場合、事前に補完クラス設定の他に、モデル選択などの時間のかかる作業が必要となるが、新調査の実施前に必要なデータが存在しない

➡ 事前に不明なデータの分布を保持できる補完方法で

■ 補完対象項目のチェック上の制約条件

- 補完対象項目が多く、項目間に合計欄とその内訳といったデータチェック上の制約がある

➡ 基本的な制約条件を満たす補完値を作りたい

結論: 比率ホットデック

6. 比率ホットデック

補完ホットデックは、通常のホットデック法の拡張で、単純なチェック制約を満たす簡便な方法

ホットデック法とは、標本データに欠測がある場合に、同じ標本の中の欠測のない値によりこれを補う方法で、標本をできるだけ均質なグループ（補完クラス）に分け、グループ内でランダムに欠測のないデータを選び補完するランダムホットデックや、何らかの基準で欠測データと最も近いものをドナーとして選択する最近隣ホットデックなどがある。

川崎・小林（1988）ホットデック法の決め付け誤差の評価方法に関する考察、統計研究彙報第46号、pp.79-95.

Memobust project team (2014) Imputation under Edit Constraints, *The Memobust Handbook on Methodology of Modern Business Statistics*.

https://ec.europa.eu/eurostat/cros/content/imputation-imputation-under-edit-constraints-pdf-file_en.

The Memobust Handbook on Methodology for Modern Business Statistics

Ad Willeboordseが編纂し、Eurostatが1997年に刊行した"Handbook on Design and Implementation of Business Surveys" 改定を目的に、2011年1月から2014年3月まで実施された「Memobust プロジェクト」の成果物。最終的には、旧版改定ではなく、フォーマットを揃えたトピック毎のPDFファイル（モジュール）の集合として全面書き換えされ、新たにThe Memobust Handbookとして刊行された。

https://ec.europa.eu/eurostat/cros/content/memobust_en

比率ホットデックの仕組み

例: 負の値をとらない変数 $y_1, \dots, y_m$ について、
 $y_1 + \dots + y_m = y_{tot}$ という制約がある場合を考える

i 番目のレコード $(y_{i,1}, \dots, y_{i,m})$ について、頭から t 個の変数が観測され、その後ろの $m-t$ 個の変数が欠測しているとき、次のような手順で補完する

- ① 欠測値の合計 $r_i = y_{i,tot} - y_{i,1} - \dots - y_{i,t}$ を計算
- ② 任意の方法で、欠測のないレコードの中からドナーを選択し、それが d 番目のレコードとする
- ③ 補完すべき変数のドナー値の合計 $r_d = y_{d,j+1} + \dots + y_{d,m}$ を計算
- ④ 下式により補完値 $\tilde{y}_{i,j}$ を計算し、補完

$$\tilde{y}_{i,j} = \frac{r_i}{r_d} y_{d,j}, \quad (j = t + 1, \dots, m)$$

ドナーは、制約を満たすチェック済データなので、同時に補完する値は自動的に制約を満たすことになる

比率ホットデックによる補完例

制約条件: 内訳1 + 内訳2 + 内訳3 = 合計

ドナーレコード d

合計

内訳 1

内訳 2

内訳 3

1,500

1000

300

200

欠測レコード i

合計

内訳 1

内訳2

内訳 3

1,200

500

?

?

$$r_d = 1500 - 1000 = 500$$

$$r_i = 1200 - 500 = 700$$

$$\tilde{y}_{i, \text{内訳}_2} = \frac{r_i}{r_d} y_{d, \text{内訳}_2} = \frac{700}{500} \times 300 = 420$$

$$\tilde{y}_{i, \text{内訳}_3} = \frac{r_i}{r_d} y_{d, \text{内訳}_3} = \frac{700}{500} \times 200 = 280$$

$$500 + 420 + 280 = 1200$$

7. 補完データの時点調整

- 新調査は31年に実施されるが、補完に使用するデータは、大部分が28年時点で、少数の29時点のものが混在する
- データ量の多いH28センサスデータと新調査データの間で、補完項目の時点間の比率を算出し、H29DBデータにはその比率の2/3を適用する
- 比率の算出は、補完クラス・補完対象項目毎に行う
- 人数データである従業者数については、時点調整は行わない

課題: 外れ値の影響緩和

8. 補完作業の流れ (1)

補完作業は、経理項目と従業者数関連項目、設備投資関連項目、専従者給与の四つに分かれ、この順番で作業を行う

■ 経理項目 [9]

ステップ A: 同一企業データを用いた補完

ステップ B: 他企業データを用いた補完

ステップ C: 経費内訳の部分的な補完

■ 従業者数関連項目 [1 0]

■ 設備投資関連項目 [1 1]

■ 専従者給与 [1 2]

8. 補完作業の流れ (2)

経理項目

05	売上金額
06	費用総額*
07	期首棚卸高
08	仕入高
09	期末棚卸高
10	経費計
11	給料賃金
12	地代家賃
13	減価償却費
14	租税公課
15	損害保険料
16	福利厚生費
17	外注工賃
18	利子割引料

9

従業者数関連項目

04	従業者数
041	事業主の家族で無給
042	常用雇用者
043	パート・アルバイト
044	臨時雇用者

10

設備投資関連項目

19	設備投資（新規設備取得）
20	うち車両 機械 工具 器具 備品
201	最も取得額の多い時期
21	設備投資（中古設備取得）

11

専従者給与

32	専従者給与
----	-------

12

9. 経理項目

A) 同一企業データからの補完

- ① 新調査の項目**07**～**10**に欠測がなければ、**07+08-09+10**により**06**を作成
- ② 項目**05**が欠測ならば、H29DBから補完する
- ③ 項目**07**～**10**に欠測があり、①で**06**を作成できない場合、H28センサスから**06**を補完
- ④ 項目**07**～**09**に1つ欠測があり、**06**が補完済みで **10**が観測されていれば、**06 - 10 = 07 + 08 - 09**という関係式に基づいて補完
- ⑤ 項目**11**～**18**が欠測の場合、H28センサスから**11**～**14**を補完するが、**10**に数値がある場合、 $\Sigma(11 \sim 14)$ がそれよりも大きい場合は補完せずに次のステップへ

制約条件: $06 + 09 = 07 + 08 + 10$
 $10 > \Sigma(11 \sim 18)$

05	売上金額
06	費用総額*
07	期首棚卸高
08	仕入高
09	期末棚卸高
10	経費計
11	給料賃金
12	地代家賃
13	減価償却費
14	租税公課
15	損害保険料
16	福利厚生費
17	外注工賃
18	利子割引料

05	売上金額
06	費用総額*
07	期首棚卸高
08	仕入高
09	期末棚卸高
10	経費計
11	給料賃金
12	地代家賃
13	減価償却費
14	租税公課
15	損害保険料
16	福利厚生費
17	外注工賃
18	利子割引料

B) 他企業データを用いた主要経理項目（06～10）の補完

ステップAの補完終了時に、欠測がないのは項目**05**のみ。

ステップBで対応すべき欠測の組み合わせは、以下の26通り存在する。

- ・ 項目07～09が1つ欠測、なおかつ06が欠測 [a～c]
- ・ 項目06と10が両方欠測 [d]
- ・ 項目07～09のみに2つ以上の欠測がある [e～g]
- ・ 項目07～09が1つ欠測、なおかつ10が欠測 [h～j]
- ・ 項目07～09が2つ以上欠測し、なおかつ06が欠測 [k～m, u]
- ・ 項目07～09が1つ欠測し、なおかつ06と10が両方欠測 [n～p]
- ・ 項目07～09がすべて欠測 [q]
- ・ 項目07～09が2つ以上欠測し、なおかつ10が欠測 [r～t]
- ・ 項目07～09が2つ以上欠測し、なおかつ06と10が両方欠測 [v～x]
- ・ 項目07～09がすべて欠測し、なおかつ10が欠測 [y]
- ・ 項目06～10がすべて欠測 [z]

ステップBで対処する欠測のパターン

05	売上金額	a. (06, 07)	n. (06, 07, 10+)
06	費用総額*	b. (06, 08)	o. (06, 08, 10+)
07	期首棚卸高	c. (06, 09)	p. (06, 09, 10+)
08	仕入高	d. (06, 10+)	q. (07, 08, 09)
09	期末棚卸高	e. (07, 08)	r. (07, 08, 10+)
10	経費計	f. (07, 09)	s. (07, 09, 10+)
11	給料賃金	g. (08, 09)	t. (08, 09, 10+)
12	地代家賃	h. (07, 10+)	u. (06, 07, 08, 09)
13	減価償却費	i. (08, 10+)	v. (06, 07, 08, 10+)
14	租税公課	j. (09, 10+)	w. (06, 07, 09, 10+)
15	損害保険料	k. (06, 07, 08)	x. (06, 08, 09, 10+)
16	福利厚生費	l. (06, 07, 09)	y. (07, 08, 09, 10+)
17	外注工賃	m. (06, 08, 09)	z. (06, 07, 08, 09, 10+)
18	利子割引料		

※ 10+は、10とその内訳である11～18の全部あるいは一部に欠測があることを示す

B) 他企業データを用いた主要経理項目 (06～10) の補完 [続き]

- 全てのパターンに比率ホットデックを使用することにより、同時に補完した範囲について **06+09=07+08+10** という制約を満たす補完値を作成する
- 制約条件の合計値**06+09**あるいは **07+08+10** を観測値で作成できる場合は、それに基づきドナーと比率を決めて、下位レベルの項目を補完する
- 欠測によりいずれの合計値（ **06+09**または **07+08+10** ）も作成できない場合、できるだけ数値の大きな項目で比率を定める

05	売上金額
06	費用総額*
07	期首棚卸高
08	仕入高
09	期末棚卸高
10	経費計
11	給料賃金
12	地代家賃
13	減価償却費
14	租税公課
15	損害保険料
16	福利厚生費
17	外注工賃
18	利子割引料

制約条件:

$$06 + 09 = 07 + 08 + 10$$

$$10 > \Sigma(11 \sim 18)$$

B) 他企業データを用いた主要経理項目 (06~10) の補完 [続き]

- **10** が欠測し、**11~14** がステップ A で補完済みの場合、**10** の補完値が **11~14** の合計値を越える値にならないければ、ドナーを廃棄し、次に近いデータを新たなドナーとする
- ドナーを替える場合、**10** と同時に補完すべき他の項目についても、新たなドナーにより補完する

※ 補完手順に複数の選択肢があるものについては、今後個人企業経済調査データでの試算により決める

※ 新調査の観測値**05**よりも、H28センサスからの補完値**06**をベースに補完した方が精度が良い可能性があり、これについても試算により決める

※ 合計値を観測値で作成できない場合の、制約条件への影響も今後検討が必要

05	売上金額
06	費用総額*
07	期首棚卸高
08	仕入高
09	期末棚卸高
10	経費計
11	給料賃金
12	地代家賃
13	減価償却費
14	租税公課
15	損害保険料
16	福利厚生費
17	外注工賃
18	利子割引料

制約条件:

$$06 + 09 = 07 + 08 + 10$$

$$10 > \Sigma(11 \sim 18)$$

具体例 (1)

05	売上金額
06	費用総額*
07	期首棚卸高
08	仕入高
09	期末棚卸高
10	経費計
11	給料賃金
12	地代家賃
13	減価償却費
14	租税公課
15	損害保険料
16	福利厚生費
17	外注工賃
18	利子割引料

合計値が作成できれば問題がない

(c) 欠測が **06, 09** で、**05, 07, 08, 10+** が観測（あるいは補完済）

同じ補完クラス内の経理項目に欠測のない新調査レコードから、補完対象レコードの **07+08+10** の値と最も近いものをドナーレコードに選択し、ドナーの各項目の値に以下の比率を乗じた値により、補完対象レコードの欠測を補完する。

$$\frac{\text{補完対象レコードの } \mathbf{07+08+10}}{\text{ドナーレコードの } \mathbf{07+08+10}}$$

チェック済みデータを用いて補完を行うため、この方法による補完は、同時に補完する項目について、制約条件を自動的に満たす

具体例 (2)

10+の具体的な対応例

9. 経理項目

05	売上金額
06	費用総額*
07	期首棚卸高
08	仕入高
09	期末棚卸高
10	経費計
11	給料賃金
12	地代家賃
13	減価償却費
14	租税公課
15	損害保険料
16	福利厚生費
17	外注工賃
18	利子割引料

(r) 欠測が **07, 08, 10+** で、**05, 06, 09**が観測（補完済）され、**10**の内訳は**11~14**が補完値、**15~18**が欠測

① 補完クラス内の欠測のないドナー候補の中から、まず **06+09** の値が欠測レコードのそれに最も近いものをドナーに選ぶ

② 次に、ドナーの各項目の値に以下の比率を乗じた値により、欠測レコードの **07, 08, 10** の補完値を作成

$$\frac{\text{補完対象レコードの } 06+09}{\text{ドナーレコードの } 06+09}$$

10 の補完値が **11~14** の合計値を越えない場合は、選んだドナーを廃棄し、**06+09** の値が次に近いものを選びなおす処理を繰り返す

③ 最終的に、**10** の補完値が②の条件を満たすとき、同一ドナーの補完値で、**07, 08, 10** の欠測を補完する

➡ **15~18**の補完は次のステップCで行う

具体例 (3)

今後検討が必要なもの

05	売上金額
06	費用総額*
07	期首棚卸高
08	仕入高
09	期末棚卸高
10	経費計
11	給料賃金
12	地代家賃
13	減価償却費
14	租税公課
15	損害保険料
16	福利厚生費
17	外注工賃
18	利子割引料

(a) 欠測が **06, 07** で、**05, 08, 09, 10+** が観測（または補完済）

- ✓ **08 + 10 - 09** の値で比率を決めて、**06**と**07**を比率ホットデックで補完
- ✓ **05** の値でドナーと比率を決めて、**06** を比率ホットデックで補完し、**06 + 09 - 08 - 10**の値を**07**とする
- ✓ **05** の値でドナーと比率を決めて、**06 + 09** を比率ホットデックで補完し、**06 + 09 - 08 - 10**の値を**07**とする

複数の案があり、どれがよいか判別困難な場合は、今後実データで比較を行い、最も良いものを採用する

C) 経費内訳の部分的な補完

10	経費計
11	給料賃金
12	地代家賃
13	減価償却費
14	租税公課
15	損害保険料
16	福利厚生費
17	外注工賃
18	利子割引料

ステップ **B** の終了時に **05～10** の欠測は全て補完される

ここでは、**10** の内訳の **11～14** が H28 センサスからの補完値で、**15～18** のみが欠測するパターンについて、**10** $- \Sigma(11 \sim 14)$ で比率を決め、**15～18** を比率ホットデックで補完する

制約条件:

$$10 > \Sigma(11 \sim 18)$$

※ ステップBにおいて、**10** は $\Sigma(11 \sim 14)$ を超える値のみを補完値としているため、ここで **10** $- \Sigma(11 \sim 14)$ がマイナスになることはない

参考資料: 別添1～2

10. 従業者数関連項目

- ① 項目**04**, **041**～**044**に欠測があれば、まずH29DBから男女別を含めて補完する
- ② ①の処理で、項目**041**～**044**の合計値に欠測が残る場合は、適切な項目を用いてドナーを選択し、項目**04**の値から項目**041**～**044**のうち観測されたデータの値を引いた数値を比率に用いる、比率ホットデックにより補完する

04	従業者数
041	事業主の家族で無給
042	常用雇用者
043	パート・アルバイト
044	臨時雇用者

制約条件:

$$04 = \Sigma(041 \sim 044)$$

11. 設備投資関連項目

19	設備投資（新規設備取得）
20	うち車両 機械 工具 器具 備品
201	最も取得額の多い時期
21	設備投資（中古設備取得）

- 設備投資があるのは、全レコードの1割前後で、売上金額等の他の経理項目のように毎年似た数値が入るわけではなく、過去の同一企業データが存在しても、その値を補完に使用するのは不適切
- 金額だけでなく、投資の有無を決める必要がある
- 項目**201** は **20** の額が最も多い四半期についてのカテゴリ項目
- 事前にデータが存在するのは、項目**19**のみ

個人企業の設備投資の割合

産業	設備投資 あり(率)
全産業	7.4%
D 建設業	4.8%
E 製造業	4.8%
G 情報通信業	15.6%
H 運輸業、郵便業	4.2%
I 卸売業、小売業	5.0%
J 金融業、保険業	7.3%
K 不動産業、物品賃貸業	6.7%
L 学術研究、専門・技術サービス業	15.1%
M 宿泊業、飲食サービス業	7.1%
N 生活関連サービス業、娯楽業	5.0%
O 教育、学習支援業	6.8%
P 医療、福祉	18.8%
Q 複合サービス事業	3.6%
R サービス業（他に分類されないもの）	6.2%

平成28年経済センサスー活動調査のうち、
平成31年個人企業経済調査の対象となる個人企業を対象として集計

設備投資の有無の推定について

使用データ: 平成24年経済センサス-活動調査

事前分析1: 設備投資の有無と他項目の相関
相関があるとみなせる項目が存在しない

事前分析2: 決定木(CART) による分析

使用変数リスト(14変数)

変数項目	変数名	変数項目	変数名
都道府県番号	K_KEN	動産・不動産賃借料	KB_URIAGE_6
開設時期	KB_KAISETSU	減価償却費	KB_URIAGE_7
開設月	KB_KAITSUKI	租税公課	KB_URIAGE_8
従業者数合計	MTX_KI_TJUGYO	設備投資の有無	KB_SETSUBI
売上金額	KB_URIAGE_1	有形固定資産	KB_YUKEI
費用総額	KB_URIAGE_2	無形固定資産	KB_MUKEI
給与総額	KB_URIAGE_4	産業大分類	KB_SANGL

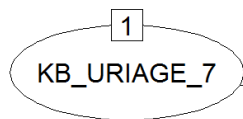
CART（分類木／回帰木）

- CARTとは、Classification And Regression Treeの略で、データマイニングや機械学習等の分野において、分類や予測のために広く使われている決定木（decision tree）と呼ばれる手法の一種
 - 目的変数がカテゴリ変数 ⇒ 分類木
 - 目的変数が連続変数 ⇒ 回帰木
- データを、グループ内の差異が小さく、グループ間の差異が大きくなるよう二分割していく
- 目的変数との関連性の高い項目ほど樹形図の上方に表示される

CARTによる樹形図

11. 設備投資関連項目

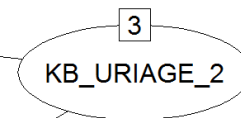
D. 建設業



≥ 48.5

< 48.5

変数項目	変数名	変数項目	変数名
都道府県番号	K_KEN	動産・不動産賃借料	KB_URIAGE_6
開設時期	KB_KAISETSU	減価償却費	KB_URIAGE_7
開設月	KB_KAITSUKI	租税公課	KB_URIAGE_8
従業者数合計	MTX_KI_TJUGYO	設備投資の有無	KB_SETSUBI
売上金額	KB_URIAGE_1	有形固定資産	KB_YUKEI
費用総額	KB_URIAGE_2	無形固定資産	KB_MUKEI
給与総額	KB_URIAGE_4	産業大分類	KB_SANGL

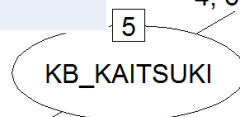


≥ 792.5

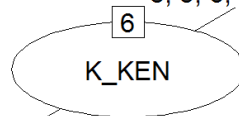
< 792.5



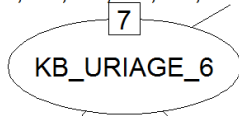
4, 6, 10, 11, 2, 3, 5, 7, 8, 9



3, 5, 6, 8, 9, 10, 1, 2, 4, 7, 11

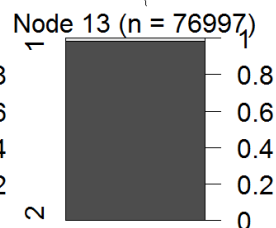
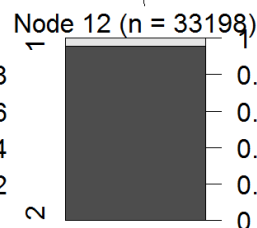
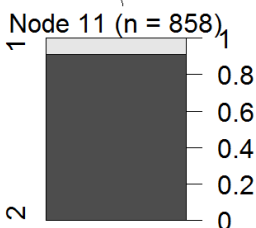
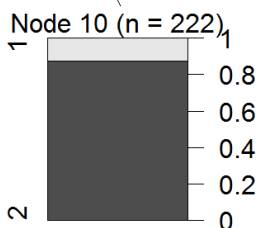
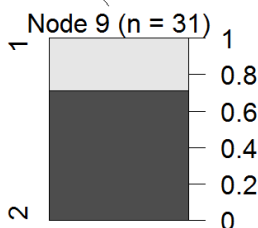
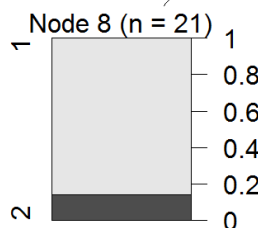
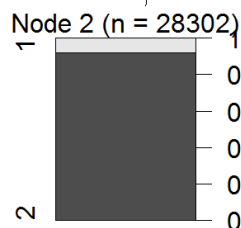


1, 5, 6, 11, 12, 13, 14, 17, 20, 23, 24, 26, 27, 30, 31, 32, 34, 36, 42, 43, 44, 45, 46



≥ 12.5

< 12.5

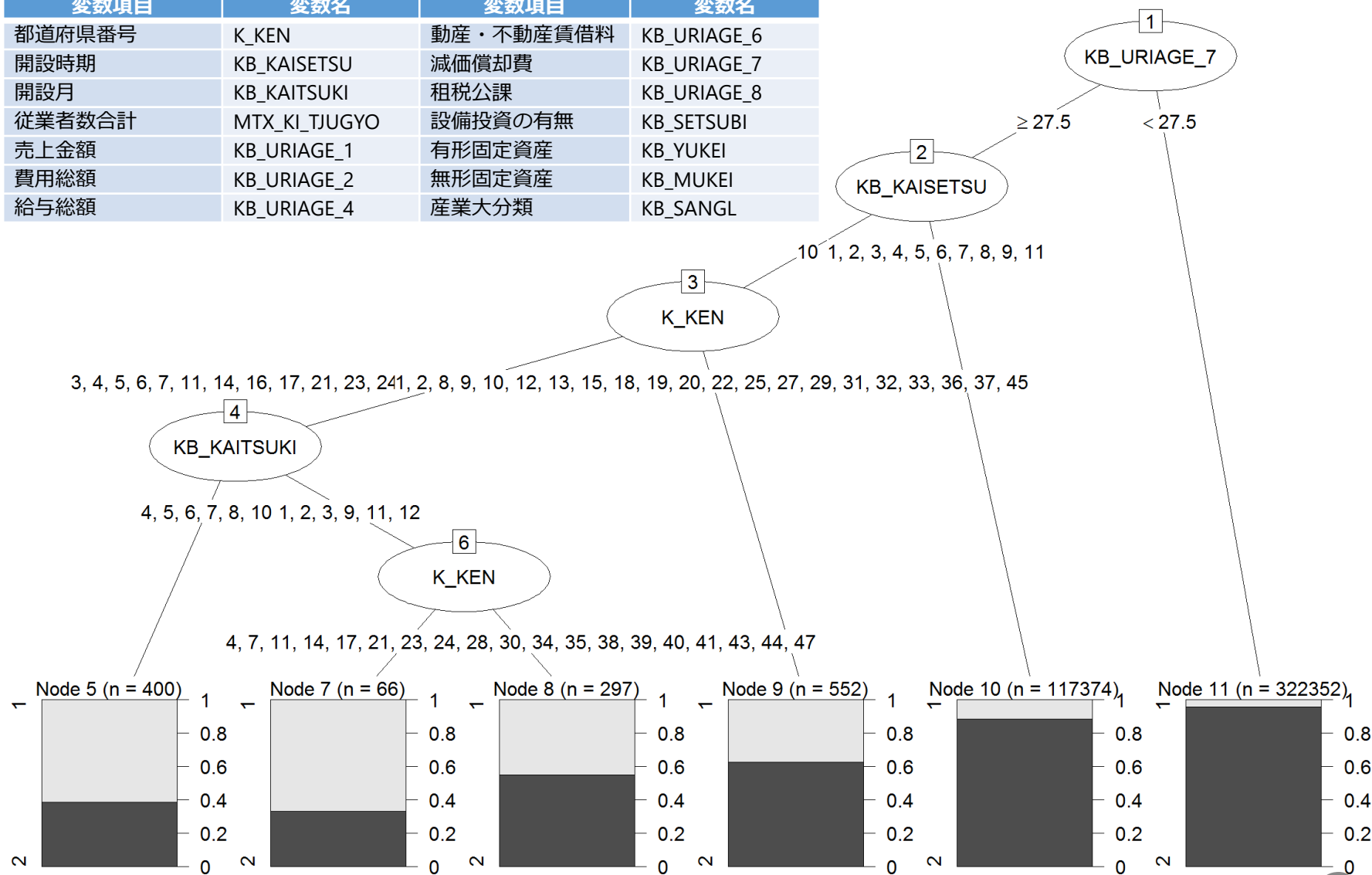


CARTによる樹形図

11. 設備投資関連項目

I. 卸売業・小売業

変数項目	変数名	変数項目	変数名
都道府県番号	K_KEN	動産・不動産賃借料	KB_URIAGE_6
開設時期	KB_KAISETSU	減価償却費	KB_URIAGE_7
開設月	KB_KAITSUKI	租税公課	KB_URIAGE_8
従業者数合計	MTX_KI_TJUGYO	設備投資の有無	KB_SETSUBI
売上金額	KB_URIAGE_1	有形固定資産	KB_YUKEI
費用総額	KB_URIAGE_2	無形固定資産	KB_MUKEI
給与総額	KB_URIAGE_4	産業大分類	KB_SANGL

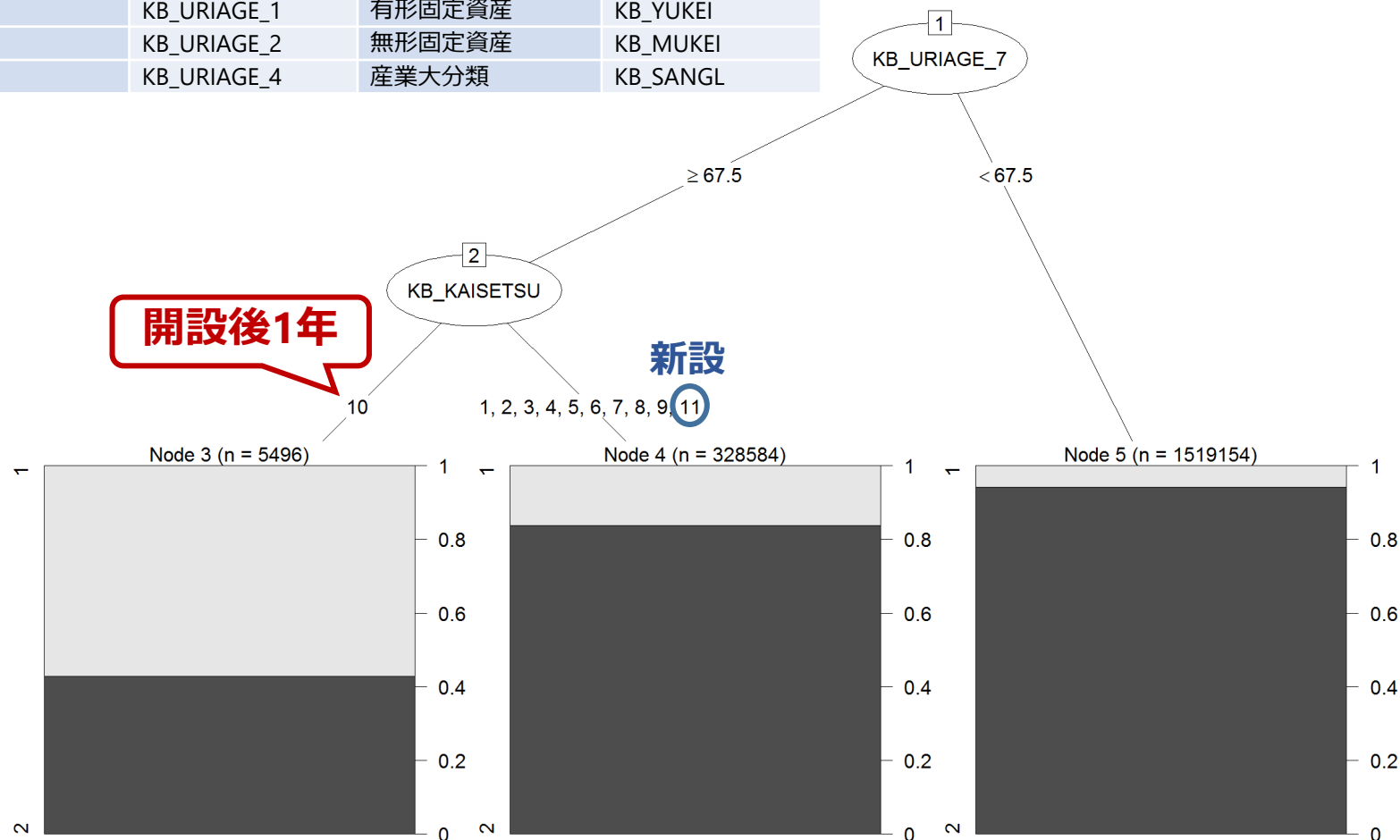


CARTによる樹形図

11. 設備投資関連項目

変数項目	変数名	変数項目	変数名
都道府県番号	K_KEN	動産・不動産賃借料	KB_URIAGE_6
開設時期	KB_KAISETSU	減価償却費	KB_URIAGE_7
開設月	KB_KAITSUKI	租税公課	KB_URIAGE_8
従業者数合計	MTX_KI_TJUGYO	設備投資の有無	KB_SETSUBI
売上金額	KB_URIAGE_1	有形固定資産	KB_YUKEI
費用総額	KB_URIAGE_2	無形固定資産	KB_MUKEI
給与総額	KB_URIAGE_4	産業大分類	KB_SANGL

全産業



CARTによる設備投資の有無の判定

- 「I. 卸売業・小売業」については、全体の9割強のデータについて、投資の有無が正しく判別されるが、量の少ない「投資あり」企業が正しく判別されるのは1%程度
- 他の産業についても、上述の傾向はほぼ同じ
- 産業によりCARTの樹形図は様々に異なる
- 多くの産業で「13 減価償却費」と「22 開設時期」が投資の有無に関係のある項目として選択され、項目22は設立後1年経過した企業に設備投資があるという結果になった
- 「I. 卸売業・小売業」のように、下位項目として県番号が選択される産業が散見される
- CARTの結果は、「投資あり」に分類される企業数が過小になる傾向がある

[19 設備投資]と他の項目との相関

使用データ: 平成24年経済センサス-活動調査

産業	従業者数						経理項目					
	個人業主	家族で無給の人	正社員	パート	臨時雇用者	従業者合計	売上金額	費用総額・経常費用	給与総額	動産・不動産賃借料	減価償却費	租税公課 (法人_住民_事業税除く)
全産業	-0.01	0.01	0.14	0.11	0.08	0.16	0.13	0.12	0.12	0.06	0.03	0.04
D	0.01	-0.07	0.14	0.02	0.05	0.05	0.13	0.16	0.06	0.11	0.15	0.06
E	0.02	-0.07	0.36	0.12	0.04	0.34	0.24	0.24	0.25	0.12	0.53	0.23
G	0.03	0.13	-0.04	0.05	0.05	0.08	0.53	0.54	0.64	0.10	0.89	0.11
H	-0.07	-0.12	-0.12	0.39	0.88	0.40	0.19	0.34	0.24	0.11	0.05	0.17
I	0.01	-0.03	0.09	0.09	0.02	0.11	0.09	0.08	0.13	0.03	0.12	0.04
J	0.11	0.18	-0.01	-0.11	0.00	-0.01	0.07	0.14	0.02	0.02	0.12	-0.06
K	-0.01	0.03	-0.02	-0.01	-0.01	-0.01	0.06	0.05	0.01	0.00	0.23	0.11
L	-0.06	0.00	0.28	0.20	0.02	0.29	0.26	0.26	0.24	0.27	0.27	0.17
M	0.01	0.10	0.12	0.16	0.04	0.19	0.12	0.02	0.14	0.01	0.00	0.11
N	0.03	-0.02	0.16	0.43	0.00	0.33	0.68	0.67	0.35	0.39	0.67	0.27
O	-0.14	-0.03	0.24	0.11	0.01	0.19	0.18	0.17	0.17	0.10	0.19	0.13
P	0.01	0.07	0.14	0.09	0.14	0.15	0.11	0.11	0.12	0.02	0.16	0.05
Q	NA	0.21	-0.02	-0.25	-0.15	-0.25	-0.63	0.26	-0.17	0.23	0.67	0.58
R	-0.01	-0.03	0.09	0.09	-0.04	0.04	0.13	0.14	0.03	0.02	0.14	0.12

※ 平成24年経済センサス-活動調査のうち、平成31年個人企業経済調査の対象となる個人企業を対象として集計

- 一部の産業には、ある程度の相関を持つ項目があるが、産業によりその項目は異なる
- ある程度の相関を持つ項目が見当たらない産業も複数存在する

設備投資関連項目の問題点

補完クラスについて

経理項目の補完において、補完クラス内でドナーとなる補完対象項目について欠測のないデータの大きさ n を確保するために、標本層の設定に採用された都道府県の別を、補完クラスの設定には使用しない。

設備投資関連項目については、投資額ゼロの企業も多く、 n の確保はさらに困難が予測される一方で、CARTの結果の下位項目で県番号が選択される産業が複数あり、地域差の存在が示唆される。

投資の有無の推定について

今後の分析で、より良い推定方法が見つかる可能性は残るが、決定木による投資の有無の推定は、投資の有無の比率を変化させる可能性があり、望ましくない。

投資額推定の問題点

項目**19**と相関が高い項目が存在するのはごく一部の産業に限られ、しかもそれは産業により異なる。また、それらの項目と、項目**20, 21**との関係は事前にはわからない。

項目19, 20及び21の補完案

案A: 投資の有無をドナーの値で決める

- [13 減価償却費] の値により、ドナーと比率を決め、項目**19, 20**及び**21**を比率ホットデックにより補完
- ドナーの投資額が0の場合は、補完値も0とする

19	設備投資（新規設備取得）
20	うち車両 機械 工具 器具 備品
201	最も取得額の多い時期
21	設備投資（中古設備取得）

案B: 投資の有無を新調査データの設備投資の有無の割合に合わせる

- 新調査データで項目**19** 及び**21** の有無の割合を算出
- 欠測のあるレコードから、算出した割合に基づきランダムに選び、項目**13**の値でドナーと比率を決め、比率ホットデックにより補完する（**19** と **20** は同時に補完する）

案Bは、投資の有無の比率は維持できるが、項目**19**と**21**に何らかの関係性がある場合にそれを維持できない

項目201の補完案

項目201はカテゴリ項目

事前に分析できるデータはH30までの個人企業動向編の対象産業のみで、記入レコード自体が非常に少ない

19	設備投資（新規設備取得）
20	うち車両 機械 工具 器具 備品
201	最も取得額の多い時期
21	設備投資（中古設備取得）

10 設備取得状況 - 昨年1年間（平成 年1月から12月まで）に 現物を取得したものについて記入してください - 耐用年数が1年以上で 取得価格が10万円以上のものについて その金額を記入してください - 分割払による購入であっても 現物を取得した時点で その総額を記入してください - 土地の購入代金は含めません - 消費税を含む金額を記入してください	(1) 新規設備取得額 億 百万 万 千 000円 (2) 中古設備取得額 億 百万 万 千 000円 ↓ うち 車両 機械 工具 器具 備品 億 百万 万 千 000円 取得額が最も多かった時期はいつですか ① 1～3月 ② 4～6月 ③ 7～9月 ④ 10～12月
--	--

- 新調査データでの四半期の別の割合に従い、ランダムに付与する
- 「不詳」符号を設定する

カテゴリデータの代表的な欠測への対応方法

12. 専従者給与について

32 専従者給与

少なくともH31新調査について、この項目の補完は、以下のような理由により推奨できない

従業者数関連項目

04	従業者数
041	事業主の家族で無給
042	常用雇用者
043	パート・アルバイト
044	臨時雇用者



- 「32 専従者給与」は、新調査実施前には利用可能なデータがない
- この項目を推定できる最有力候補は、専従者の人数であるが、新調査前も後も、当該データは存在しない
- 既存項目で専従者の人数に一番近い候補は「042 常用雇用者」になるが、事前に推定精度は確認できない

13. 今後の作業と課題

- H30.07から利用可能になったH28センサスデータによる補完方法の検証作業を行う
- センサスあるいは母集団情報の当該企業データから補完する場合、時点の違いの調整を行うが、この際に外れ値の影響緩和を試みる
- 比率ホットデックを行う際に、極端なレコードがドナーとならないように、前処理として多変量外れ値検出法の適用を検討する
- 補完クラス設定は従来CARTで行っているが、より結果の安定性に優れるルールアンサンブル法の適用もできれば検討したい
- 単位欠測について、補完をせず乗率調整のみによる集計や、単位欠測データの乗率を1に修正した集計も行った方がよい