別紙 3

総合テスト及びドナー候補に対する外れ値検出の閾値の検討

新調査の経理項目の欠測値補完処理において、欠測する項目に過去の同一企業データが存在する場合は、まずその過去の情報により補完を行うが、過去データについては調査年が異なるため、時点調整をする。時点調整が必要な過去データが存在する項目は、売上金額、費用総額と経費計、経費計の内訳である給与賃金、地代家賃、減価償却費、租税公課である。

時点調整を行った過去データによる補完の後に、同一企業データが過去に存在しない項目について、新調査の他企業データを用いた比率ホットデック補完を行う。その際に補完値として特異なドナーが選択されることを防ぐため、ドナー候補データに対して多変量外れ値検出を行う。

ここでは、経済統計課からの依頼により、実際の補完処理に即して人為的に欠測を導入したデータに対し、比率ホットデック補完だけでなく、時点調整及び外れ値処理を加えた総合的な補完のシミュレーションを行い、最後に外れ値検出の閾値を検討した。

使用データ

ここでは、新調査の経理項目をすべて含む平成 30 年の個人企業データ（データサイズ 3,585）のうち H28 経済センサスデータと事業所キーにより企業マッチングができたデータ（データサイズ 2,770）を使用する。

時点調整可能な産業別、売上高階級別の企業数および欠測率別の企業数を表 1 に示す。

表 1　時点調整シミュレーション可能な産業、売上高階級、欠測率別企業数

産業	売上高階級	全データ	欠測率		
			20%欠測	30%欠測	40%欠測
E	90%以上	45	9	14	18
I		88	18	26	35
M		63	13	19	25
Q		69	14	21	28
E	90%未満	364	73	109	146
I		903	181	271	361
M		574	115	172	230
Q		664	133	199	266

補完処理を行う単位となる補完クラスは、産業、売上高階級別に設定するため、本来ここでの補完シミュレーションは、同じクラスで検証を行う必要がある。しかし、時点調整が可能なデータが少なく、欠測率が高い場合にドナー候補の数が過小になるため、同一ドナーが繰り返し選択される可能性がある。これを防ぐため、今回のシミュレーションでは売上高階級を分けずに産業別に補完クラスを設定した。

また、時点調整に使用する比率は、表 2 に示す平成 30 年データの全産業一括で算出した比率を使用する。

表 2　時点調整の項目別の比率

項目	売上金額	費用総額
比率	0.93	1.09

ドナー候補に対する外れ値の除外処理

時点調整が可能なレコードから 20%から 50%まで任意に設定した欠測率に応じてランダムに欠測とみなすデータを選択し、それ以外の欠測とみなさないレコードをドナーデータとして、任意に設定した閾値別に外れ値検出を行った。

外れ値として検出されるレコード数を産業別、欠測率別に表 3 に示す。データの欠測率が上がればドナー候補のデータの数が減少するため、外れ値と判定されるデータ数も一般には減少傾向となるが、産業により分散が大きくなった結果、外れ値と判定されるデータ数が増加するものもある。また、外れ値判定の閾値を上げれば、当然ながら外れ値として検出されるデータ数は減少していく。

閾値を F 検定統計量の 99.9%点（つまりデータが F 分布に従うと仮定した場合に、全体の 99.9%の観測値が収まる範囲）の 1.2 倍に設定する場合、欠測率 20%で全産業では 5.7%のデータが外れ値として検出された。欠測率を 30%では、これが全産業で 5.4%と減少する。また、閾値を 1.5 倍にした場合、欠測率 20%では、全産業で 4%のデータが外れ値として検出され、欠測率 30%では、全産業で 3.7%となった。産業別にみると、データサイズの小さい産業 E（製造業）は、分散も小さいために外れ値検出されるデータ数が少ない傾向にある。産業別に欠測率 30%のドナーデータについて、閾値を 1.5 とした外れ値分布を図 1 ～ 4 に示す。外れ値データを赤丸で示している。多変量の外れ値検出法を使用しているため、データ全体の分布の中心部から外れたデータが外れ値と判定されるが、個々の項目の値が極端に大きい場合と、極端に大きい項目がなくても、相関関係が他の大部分のデータと大きく異なる場合がある。

表 3　閾値別、欠測率別の外れ値データ数及び割合

閾値	産業	データ数	マッチングデータ数	20%欠測				30%欠測				40%欠測				50%欠測			
				欠測データ	ドナーデータ	外れ値データ	(%)	欠測データ	ドナーデータ	外れ値データ	(%)	欠測データ	ドナーデータ	外れ値データ	(%)	欠測データ	ドナーデータ	外れ値データ	(%)
1	E	515	409	103	412	20	4.9%	154	361	15	4.2%	206	309	14	4.5%	257	258	10	3.9%
	I	1292	991	258	1034	86	8.3%	387	905	75	8.3%	516	776	62	8.0%	646	646	55	8.5%
	M	815	637	163	652	43	6.6%	244	571	38	6.7%	326	489	37	7.6%	407	408	26	6.4%
	Q	963	733	192	771	66	8.6%	288	675	53	7.9%	385	578	46	8.0%	481	482	37	7.7%
	全産業	3585	2770	716	2869	215	7.5%	1073	2512	181	7.2%	1433	2152	159	7.4%	1791	1794	128	7.1%
1.2	E	515	409	103	412	17	4.1%	154	361	13	3.6%	206	309	10	3.2%	257	258	9	3.5%
	I	1292	991	258	1034	61	5.9%	387	905	53	5.9%	516	776	44	5.7%	646	646	35	5.4%
	M	815	637	163	652	36	5.5%	244	571	30	5.3%	326	489	27	5.5%	407	408	20	4.9%
	Q	963	733	192	771	50	6.5%	288	675	39	5.8%	385	578	34	5.9%	481	482	26	5.4%
	全産業	3585	2770	716	2869	164	5.7%	1073	2512	135	5.4%	1433	2152	115	5.3%	1791	1794	90	5.0%
1.3	E	515	409	103	412	13	3.2%	154	361	10	2.8%	206	309	8	2.6%	257	258	5	1.9%
	I	1292	991	258	1034	50	4.8%	387	905	39	4.3%	516	776	37	4.8%	646	646	31	4.8%
	M	815	637	163	652	35	5.4%	244	571	30	5.3%	326	489	26	5.3%	407	408	19	4.7%
	Q	963	733	192	771	46	6.0%	288	675	38	5.6%	385	578	30	5.2%	481	482	24	5.0%
	全産業	3585	2770	716	2869	144	5.0%	1073	2512	117	4.7%	1433	2152	101	4.7%	1791	1794	79	4.4%
1.5	E	515	409	103	412	9	2.2%	154	361	8	2.2%	206	309	6	1.9%	257	258	3	1.2%
	I	1292	991	258	1034	36	3.5%	387	905	29	3.2%	516	776	26	3.4%	646	646	21	3.3%
	M	815	637	163	652	34	5.2%	244	571	28	4.9%	326	489	24	4.9%	407	408	18	4.4%
	Q	963	733	192	771	37	4.8%	288	675	28	4.1%	385	578	25	4.3%	481	482	19	3.9%
	全産業	3585	2770	716	2869	116	4.0%	1073	2512	93	3.7%	1433	2152	81	3.8%	1791	1794	61	3.4%
2	E	515	409	103	412	8	1.9%	154	361	8	2.2%	206	309	6	1.9%	257	258	2	0.8%
	I	1292	991	258	1034	20	1.9%	387	905	17	1.9%	516	776	13	1.7%	646	646	9	1.4%
	M	815	637	163	652	27	4.1%	244	571	24	4.2%	326	489	21	4.3%	407	408	15	3.7%
	Q	963	733	192	771	23	3.0%	288	675	17	2.5%	385	578	14	2.4%	481	482	12	2.5%
	全産業	3585	2770	716	2869	78	2.7%	1073	2512	66	2.6%	1433	2152	54	2.5%	1791	1794	38	2.1%
3	E	515	409	103	412	6	1.5%	154	361	6	1.7%	206	309	4	1.3%	257	258	1	0.4%
	I	1292	991	258	1034	15	1.5%	387	905	13	1.4%	516	776	10	1.3%	646	646	7	1.1%
	M	815	637	163	652	18	2.8%	244	571	16	2.8%	326	489	15	3.1%	407	408	10	2.5%
	Q	963	733	192	771	15	1.9%	288	675	10	1.5%	385	578	8	1.4%	481	482	8	1.7%
	全産業	3585	2770	716	2869	54	1.9%	1073	2512	45	1.8%	1433	2152	37	1.7%	1791	1794	26	1.4%
5	E	515	409	103	412	1	0.2%	154	361	1	0.3%	206	309	1	0.3%	257	258	1	0.4%
	I	1292	991	258	1034	10	1.0%	387	905	9	1.0%	516	776	6	0.8%	646	646	4	0.6%
	M	815	637	163	652	11	1.7%	244	571	11	1.9%	326	489	10	2.0%	407	408	7	1.7%
	Q	963	733	192	771	8	1.0%	288	675	7	1.0%	385	578	5	0.9%	481	482	4	0.8%
	全産業	3585	2770	716	2869	30	1.0%	1073	2512	28	1.1%	1433	2152	22	1.0%	1791	1794	16	0.9%

図 1　産業 E 欠測率 30%のドナーデータ分布（赤は外れ値で閾値の乗数は 1.5）

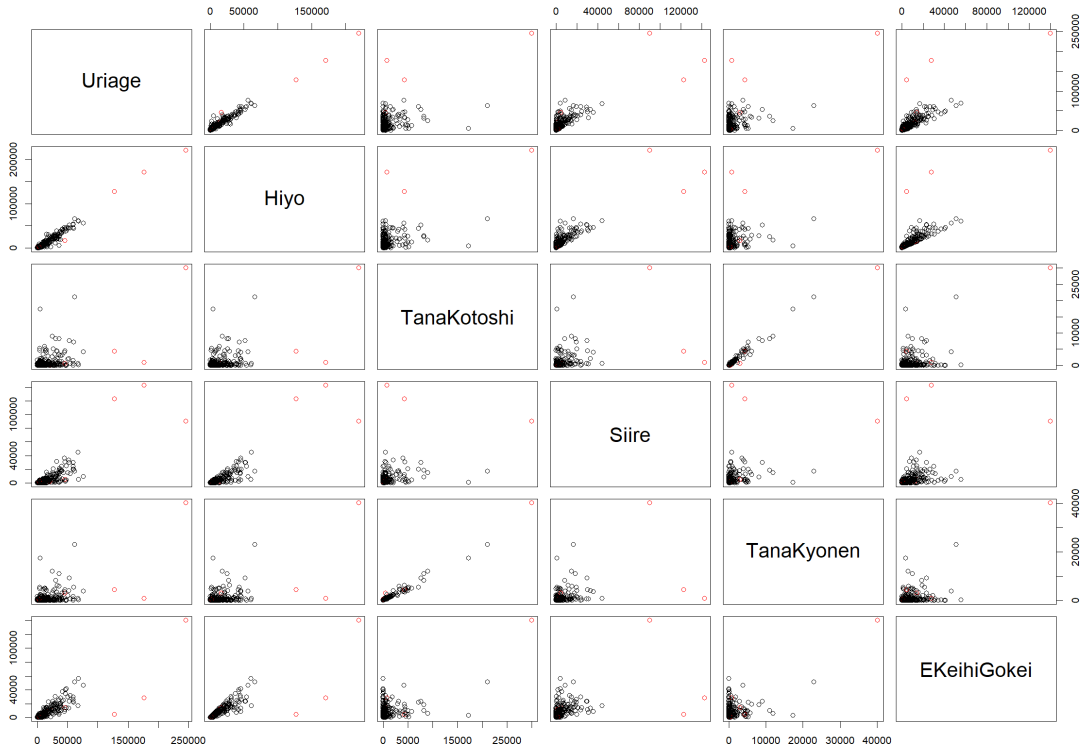


図 2　産業 I 欠測率 30%のドナーデータ分布（赤は外れ値で閾値の乗数は 1.5）

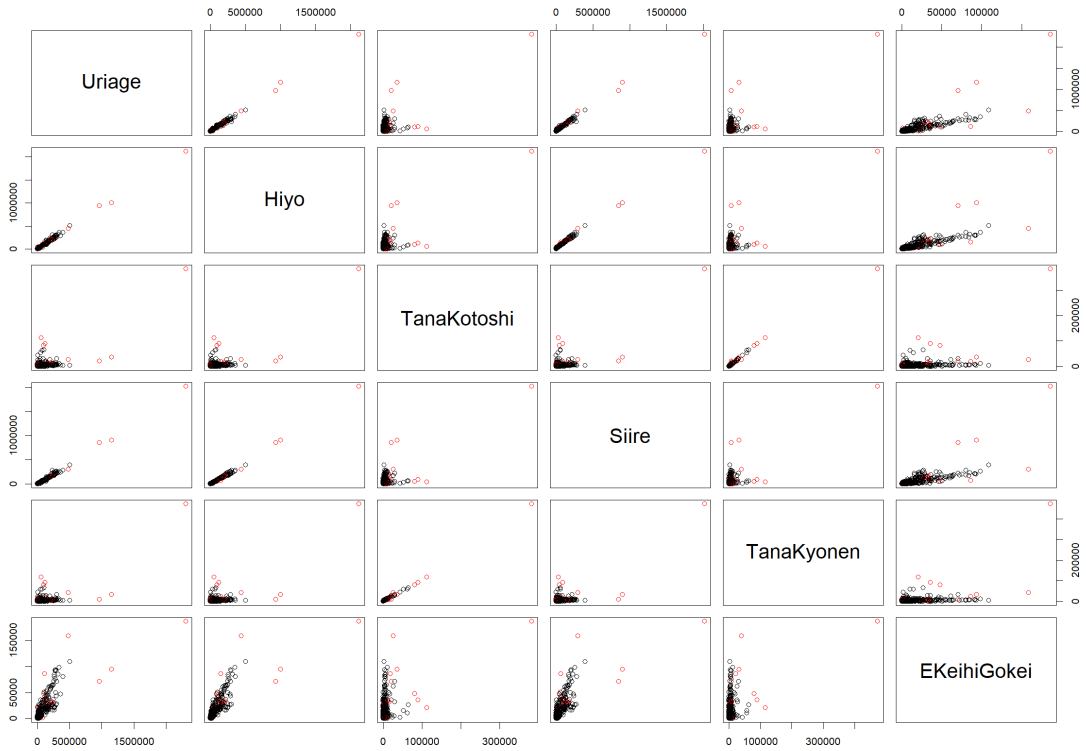


図 3　産業 M 欠測率 30%のドナーデータ分布（赤は外れ値で閾値の乗数は 1.5）

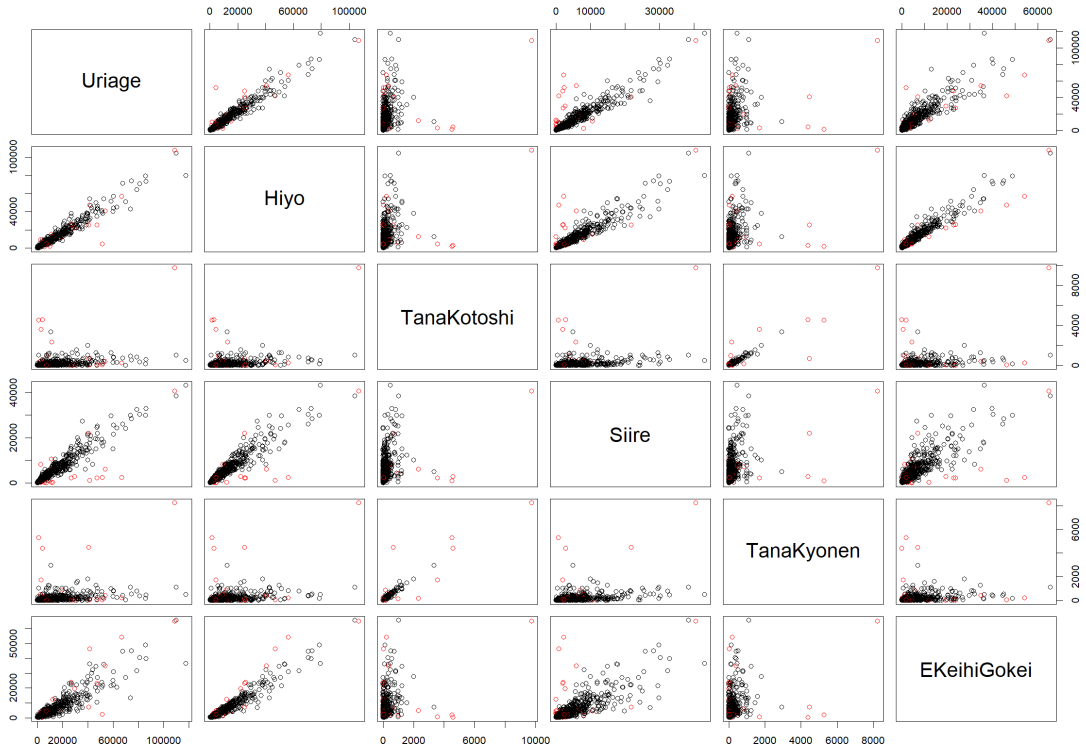
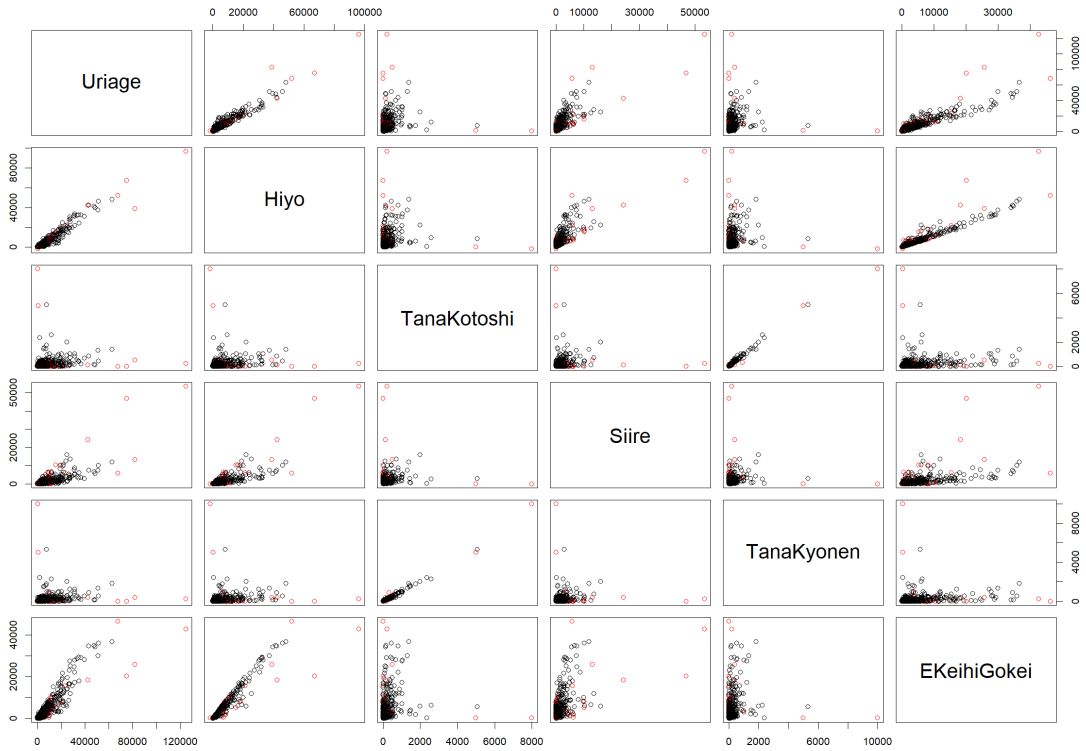


図 4　産業 Q 欠測率 30%のドナーデータ分布（赤は外れ値で閾値の乗数は 1.5）



ここまでのシミュレーションは、ある程度のドナー候補の大きさを確保するために、産業別に補完クラスを設定したが、さらに閾値について検討するために、実際の新調査集計時と同様に、産業別及び売上高階級別で補完クラスを設定して外れ値処理を行い、その結果を表4に示す。この場合、売上高階級を分けずに産業別に行った場合よりも外れ値が増加傾向にある。また、売上高90%以上と90%未満ではデータ数が異なるが、外れ値として検出されるデータ数の割合は、売上高階級90%以上の方が多い傾向にあり、これは、売上高階級90%以上の補完クラスのデータ分散が大きいことに起因する。

表 4　閾値別、欠測率、売上高階級別の外れ値データ数

閾値	産業	データ数	売上高階級別外れ値データ数															
			20%欠測				30%欠測				40%欠測				50%欠測			
			90%以上		90%未満		90%以上		90%未満		90%以上		90%未満		90%以上		90%未満	
ドナー数	E製造業		42		371		37		325		32		278		26		232	
	I卸売業、小売業		104		931		91		815		78		698		65		582	
	M宿泊業、飲食サービス業		68		585		59		512		51		439		42		366	
	Qサービス業		78		693		68		607		59		520		49		433	
1	E	515	6	14.3%	16	4.3%	6	16.2%	16	4.9%	6	18.8%	13	4.7%	5	19.2%	9	3.9%
	I	1292	27	26.0%	65	7.0%	24	26.4%	54	6.6%	19	24.4%	46	6.6%	17	26.2%	40	6.9%
	M	815	7	10.3%	34	5.8%	7	11.9%	29	5.7%	5	9.8%	24	5.5%	4	9.5%	19	5.2%
	Q	963	9	11.5%	55	7.9%	8	11.8%	50	8.2%	7	11.9%	41	7.9%	7	14.3%	35	8.1%
1.2	E	515	6	14.3%	9	2.4%	6	16.2%	10	3.1%	6	18.8%	8	2.9%	5	19.2%	5	2.2%
	I	1292	24	23.1%	46	4.9%	22	24.2%	42	5.2%	17	21.8%	35	5.0%	13	20.0%	31	5.3%
	M	815	5	7.4%	26	4.4%	5	8.5%	21	4.1%	5	9.8%	17	3.9%	4	9.5%	13	3.6%
	Q	963	8	10.3%	41	5.9%	7	10.3%	38	6.3%	6	10.2%	30	5.8%	6	12.2%	25	5.8%
1.3	E	515	6	14.3%	9	2.4%	6	16.2%	9	2.8%	6	18.8%	8	2.9%	5	19.2%	5	2.2%
	I	1292	23	22.1%	43	4.6%	20	22.0%	38	4.7%	17	21.8%	33	4.7%	12	18.5%	29	5.0%
	M	815	5	7.4%	24	4.1%	5	8.5%	19	3.7%	4	7.8%	16	3.6%	4	9.5%	12	3.3%
	Q	963	8	10.3%	35	5.1%	7	10.3%	31	5.1%	5	8.5%	23	4.4%	6	12.2%	18	4.2%
1.5	E	515	6	14.3%	6	1.6%	6	16.2%	7	2.2%	6	18.8%	6	2.2%	5	19.2%	5	2.2%
	I	1292	19	18.3%	29	3.1%	17	18.7%	25	3.1%	16	20.5%	18	2.6%	12	18.5%	15	2.6%
	M	815	5	7.4%	24	4.1%	5	8.5%	19	3.7%	4	7.8%	15	3.4%	3	7.1%	12	3.3%
	Q	963	6	7.7%	28	4.0%	6	8.8%	27	4.4%	3	5.1%	21	4.0%	5	10.2%	17	3.9%
2	E	515	6	14.3%	5	1.3%	6	16.2%	5	1.5%	4	12.5%	5	1.8%	3	11.5%	4	1.7%
	I	1292	19	18.3%	18	1.9%	17	18.7%	16	2.0%	13	16.7%	13	1.9%	12	18.5%	11	1.9%
	M	815	5	7.4%	21	3.6%	5	8.5%	16	3.1%	4	7.8%	15	3.4%	3	7.1%	11	3.0%
	Q	963	3	3.8%	20	2.9%	1	1.5%	19	3.1%	1	1.7%	14	2.7%	0	0.0%	11	2.5%
3	E	515	4	9.5%	4	1.1%	3	8.1%	4	1.2%	3	9.4%	4	1.4%	2	7.7%	1	0.4%
	I	1292	14	13.5%	14	1.5%	13	14.3%	13	1.6%	11	14.1%	10	1.4%	8	12.3%	8	1.4%
	M	815	4	5.9%	12	2.1%	4	6.8%	8	1.6%	3	5.9%	7	1.6%	2	4.8%	4	1.1%
	Q	963	1	1.3%	14	2.0%	1	1.5%	13	2.1%	1	1.7%	10	1.9%	0	0.0%	9	2.1%
5	E	515	3	7.1%	1	0.3%	1	2.7%	1	0.3%	0	0.0%	1	0.4%	1	3.8%	1	0.4%
	I	1292	12	11.5%	6	0.6%	11	12.1%	5	0.6%	10	12.8%	4	0.6%	6	9.2%	4	0.7%
	M	815	0	0.0%	7	1.2%	0	0.0%	4	0.8%	1	2.0%	5	1.1%	1	2.4%	3	0.8%
	Q	963	0	0.0%	10	1.4%	0	0.0%	10	1.6%	0	0.0%	7	1.3%	0	0.0%	5	1.2%

比率ホットデック補完シミュレーションのパターン

比率ホットデック補完を行う際に、売上金額と費用総額については、まず過去の同一企業データが存在する場合は時点調整を施した上で補完値として使用する。ここでは、こられの項目について、時点調整して補完した場合と、時点調整せず欠測としたまま比率ホットデック補完を行った場合について比較を行う。

さらに、比率ホットデック補完においては、ドナーレコードに対して外れ値処理を行う場合と行わない場合について比較する。

ここで検証を行った比率ホットデックのパターンを、表 5 に示す。

表 5　シミュレーションパターン

No	時点調整		外れ値処理	パターン
	売上金額	費用総額		
①	-	×	×	k,l,m,n,o,p,u,v,w,x,z
②	-	×	○	k,l,m,n,o,p,u,v,w,x,z
③	-	○	×	e,f,g,h,l,j,q,r,s,t,y
④	-	○	○	e,f,g,h,l,j,q,r,s,t,y
⑤	○	×	×	k,l,m,n,o,p,u,v,w,x,z
⑥	○	×	○	k,l,m,n,o,p,u,v,w,x,z

表 5 において、時点調整の売上金額及び費用総額が「－」の場合は、その項目が観測されたことを示し、「○」は時点調整を行った過去データにより補完する場合を、「×」は過去データを使用せずに欠測のままとして比率ホットデックにより補完を行うことを示している。なお、売上金額及び費用総額の両方が、時点調整を行った過去データによる補完値である場合に、後に続く比率ホットデック補完においては売上金額の補完値を使用せず、③及び④のパターンと結果は同一となる。①と②のパターン k～p、u～x、z については、③と④のパターン e～j、q～t、y の場合の複数の欠測項目に費用総額が含まれているが、過去の同一企業の費用総額データが存在しないために、費用総額についても比率ホットデックを必要としている。

このシミュレーションにおいて外れ値処理で使う閾値は 99.9%値の 1.5 倍を採用した。

シミュレーションの結果

比率ホットデック補完のシミュレーション同様、①～⑥について、真値と補完値との誤差を欠測率 20%で 100 回のシミュレーションを行った平均の NRMSE を表６～8 に示す。

欠測率 20%で外れ値の閾値 1.5 倍の場合について、費用総額を売上金額から補完する①もしくは②と、過去の同一企業データから時点調整を行って費用総額を補完する③もしくは④の結果を比較すると、同一企業の費用総額を時点調整するよりも、売上金額から補完する①②の方が真値と補完値の誤差は小さくなる傾向にある。特に費用総額は売上金額の近いドナーから比率ホットデック補完する結果の方が、誤差がより小さくなる傾向がみられた。また、外れ値処理の有無については、わずかに誤差が小さくなるパターンや産業及び項目があったが、誤差が増加傾向にある項目も存在する。元々、期首棚卸高及び期末棚卸高の両方が欠測する場合は特に、両者と相関が高い他の項目は存在しないために良いドナーの選択が困難であるという問題がある。

比率ホットデックを行う際に、ドナーレコードへの外れ値処理を行わない場合と行う場合を比較すると、わずかではあるが外れ値処理を行った場合の方が NRMSE は 0 に近づき、真値と補完値との誤差が小さくなる傾向が見られた。しかし、産業や欠測パターンによりあまり変わらない、または誤差が大きくなる箇所も散見される。

いずれのパターンにおいても、欠測率が上がることで NRMSE は大きくなる傾向にあり、欠測率が上がればドナーデータの数が減少し、結果としてデータ分散が大きくなって検出される外れ値も増えてさらにデータ数が減少し、これがある程度以上少なくなれば、適切なドナーレコードが近隣に存在しないために、真値と補完値との誤差が大きくなることが考えられる。これを防ぐためには、補完クラスはあまり小さくしない設定が望ましいということが示唆される。

表 6　①②の結果

パターン	産業	外れ値なし					外れ値処理あり				
		費用総額	棚卸期首	仕入金額	棚卸期末	経費計	費用総額	棚卸期首	仕入金額	棚卸期末	経費計
k	E	0.34	0.63	0.54			0.35	0.40	0.54		
	I	0.16	1.60	0.41			0.16	1.19	0.30		
	M	0.42	0.70	0.77			0.38	0.91	0.72		
	Q	0.46	0.78	0.99			0.45	0.62	0.92		
l	E	0.34	6.49		4.58		0.35	7.16		5.18	
	I	0.16	2.83		2.29		0.16	12.71		10.89	
	M	0.41	18.16		12.97		0.38	43.01		38.91	
	Q	0.46	29.52		24.96		0.46	31.81		27.10	
m	E	0.34		0.92	2.20		0.35		0.77	1.78	
	I	0.16		0.88	5.47		0.16		9.09	42.70	
	M	0.41		0.80	1.17		0.38		0.77	1.38	
	Q	0.46		1.00	1.10		0.45		0.98	1.03	
n	E	0.34	0.58			0.61	0.35	0.43			0.61
	I	0.16	0.43			0.69	0.16	0.42			0.65
	M	0.41	0.91			0.70	0.38	0.76			0.66
	Q	0.46	0.54			0.66	0.45	0.60			0.63
o	E	0.34		0.70		0.77	0.35		0.74		0.75
	I	0.16		0.26		1.00	0.16		0.30		1.54
	M	0.42		0.62		0.70	0.38		0.51		0.61
	Q	0.46		0.96		0.69	0.45		0.87		0.70
p	E	0.34			8.58	2.28	0.35			1.74	0.76
	I	0.16			2.70	1.47	0.16			3.58	1.84
	M	0.42			0.94	0.70	0.38			1.24	0.66
	Q	0.46			2.00	0.71	0.45			2.31	0.74
u	E	0.34	1.33	0.51	1.36		0.35	1.31	0.54	1.35	
	I	0.16	1.37	0.20	1.36		0.16	1.24	0.19	1.24	
	M	0.41	1.33	0.77	1.32		0.38	1.10	0.73	1.10	
	Q	0.46	1.98	0.92	1.82		0.45	1.43	0.90	1.46	
v	E	0.34	0.48	0.65		0.75	0.35	0.40	0.70		0.73
	I	0.16	1.47	0.42		0.67	0.16	1.07	0.33		0.87
	M	0.41	0.90	0.60		0.71	0.38	0.90	0.50		0.60
	Q	0.46	0.95	1.08		0.70	0.45	0.59	0.86		0.68
w	E	0.34	1.46		1.57	0.58	0.35	0.96		1.10	0.60
	I	0.16	2.19		2.13	0.75	0.16	2.30		2.17	0.70
	M	0.41	7.33		7.90	0.71	0.38	0.89		0.91	0.66
	Q	0.46	4.51		4.52	0.64	0.45	2.78		2.96	0.64
x	E	0.34		3.03	16.03	1.28	0.35		3.06	16.04	1.25
	I	0.16		0.52	3.59	2.25	0.16		0.65	4.41	2.89
	M	0.42		0.63	1.82	0.70	0.38		0.49	1.63	0.62
	Q	0.46		1.01	0.85	0.70	0.45		0.87	0.88	0.71
z	E	0.34	1.52	0.76	1.55	0.84	0.35	1.28	0.88	1.30	0.93
	I	0.16	1.51	0.21	1.52	0.68	0.16	1.35	0.22	1.33	0.90
	M	0.41	1.48	0.51	1.35	0.71	0.38	1.09	0.43	1.08	0.70
	Q	0.46	1.43	1.01	1.47	0.69	0.45	1.22	0.84	1.25	0.68

表 7　③④の結果

パターン	産業	外れ値処理なし					外れ値処理あり				
		費用総額	棚卸期首	仕入金額	棚卸期末	経費計	費用総額	棚卸期首	仕入金額	棚卸期末	経費計
e	E	0.57	0.69	0.85			0.57	0.63	0.84		
	I	0.52	1.51	0.73			0.52	1.36	0.66		
	M	0.48	0.90	0.90			0.48	0.91	0.89		
	Q	0.59	0.61	1.24			0.59	0.63	1.21		
f	E	0.57	11.35		7.76		0.57	12.84		9.18	
	I	0.52	5.75		2.44		0.52	49.41		42.10	
	M	0.48	22.65		16.09		0.48	55.96		50.32	
	Q	0.59	38.83		32.92		0.59	42.26		36.07	
g	E	0.57		1.50	4.23		0.57		1.24	2.85	
	I	0.52		1.13	5.19		0.52		11.21	49.71	
	M	0.48		0.95	1.32		0.48		0.98	2.05	
	Q	0.59		1.25	1.08		0.59		1.25	0.97	
h	E	0.57	0.62			1.07	0.57	0.39			1.07
	I	0.52	1.10			1.69	0.52	0.79			1.80
	M	0.48	0.76			0.83	0.48	0.86			0.80
	Q	0.59	0.72			0.79	0.59	0.63			0.79
i	E	0.57		0.85		0.79	0.57		0.79		0.86
	I	0.52		0.57		1.21	0.52		0.58		1.84
	M	0.48		0.73		0.72	0.48		0.63		0.65
	Q	0.59		0.94		0.76	0.59		0.93		0.74
j	E	0.57			8.42	2.68	0.57			1.17	1.09
	I	0.52			16.41	26.90	0.52			15.89	26.71
	M	0.48			1.09	0.83	0.48			1.68	0.84
	Q	0.59			1.74	0.82	0.59			2.05	0.82
q	E	0.57	1.28	0.83	1.34		0.57	1.26	0.83	1.32	
	I	0.52	1.50	0.61	1.49		0.52	1.53	0.62	1.54	
	M	0.48	1.44	0.90	1.40		0.48	1.08	0.90	1.09	
	Q	0.59	1.56	1.24	1.53		0.59	1.43	1.24	1.44	
r	E	0.57	0.51	0.82		0.81	0.57	0.36	0.78		0.84
	I	0.52	1.82	0.72		0.77	0.52	1.21	0.61		1.02
	M	0.48	1.03	0.71		0.78	0.48	0.91	0.65		0.66
	Q	0.59	0.72	0.96		0.76	0.59	0.51	0.95		0.74
s	E	0.57	2.03		2.11	1.06	0.57	0.93		1.01	1.05
	I	0.52	5.78		5.03	2.04	0.52	2.95		2.74	2.06
	M	0.48	3.89		4.07	0.82	0.48	1.14		1.15	0.81
	Q	0.59	4.32		4.38	0.79	0.59	1.83		1.99	0.79
t	E	0.57		0.87	0.78	0.84	0.57		0.85	1.44	0.86
	I	0.52		1.06	4.47	2.95	0.52		1.09	5.12	3.49
	M	0.48		0.77	2.61	0.74	0.48		0.68	2.66	0.69
	Q	0.59		0.93	0.67	0.75	0.59		0.97	0.65	0.75
y	E	0.57	1.80	0.77	1.95	1.18	0.57	2.20	0.75	2.30	1.36
	I	0.52	1.30	0.57	1.30	0.72	0.52	1.28	0.52	1.26	1.09
	M	0.48	1.22	0.70	1.28	0.66	0.48	1.10	0.67	1.08	0.66
	Q	0.59	1.07	0.95	1.11	0.72	0.59	1.08	0.96	1.12	0.72

表 8　⑤⑥の結果

パターン	産業	外れ値処理なし					外れ値処理あり				
		費用総額	棚卸期首	仕入金額	棚卸期末	経費計	費用総額	棚卸期首	仕入金額	棚卸期末	経費計
k	E	0.47	0.70	0.72			0.46	0.49	0.69		
	I	0.57	1.52	0.87			0.58	1.16	0.80		
	M	0.47	0.76	0.88			0.47	0.87	0.89		
	Q	0.58	0.83	1.23			0.57	0.71	1.14		
l	E	0.48	8.94		6.21		0.46	8.94		6.47	
	I	0.57	7.64		2.70		0.58	67.39		57.80	
	M	0.47	21.99		15.50		0.47	55.31		49.29	
	Q	0.58	37.46		31.79		0.57	39.04		33.35	
m	E	0.48		0.89	1.54		0.46		0.79	1.39	
	I	0.57		1.30	6.40		0.58		8.93	40.74	
	M	0.47		0.91	1.26		0.47		0.91	1.25	
	Q	0.58		1.22	1.17		0.57		1.17	1.12	
n	E	0.47	0.50			0.89	0.46	0.45			0.79
	I	0.57	1.04			2.19	0.58	0.70			2.41
	M	0.47	0.88			0.80	0.47	0.74			0.80
	Q	0.58	0.50			0.81	0.57	0.58			0.79
o	E	0.47		0.72		0.74	0.46		0.70		0.76
	I	0.57		0.71		0.96	0.58		0.71		1.62
	M	0.47		0.64		0.76	0.47		0.62		0.68
	Q	0.58		1.00		0.78	0.57		0.90		0.77
p	E	0.47			5.03	1.65	0.46			1.62	0.91
	I	0.57			5.86	4.37	0.58			6.31	4.60
	M	0.47			1.60	0.81	0.47			1.19	0.81
	Q	0.58			2.62	0.89	0.57			4.40	1.02
u	E	0.47	1.26	0.70	1.29		0.46		0.68	1.27	
	I	0.57	2.07	0.73	2.19		0.58		0.72	1.41	
	M	0.47	1.48	0.89	1.35		0.47		0.90	1.06	
	Q	0.58	2.23	1.18	2.12		0.57		1.13	1.52	
v	E	0.47	0.50	0.69		0.77	0.46	0.40	0.70		0.74
	I	0.57	1.41	0.84		0.72	0.58	0.99	0.73		1.04
	M	0.47	0.84	0.60		0.78	0.47	0.86	0.59		0.68
	Q	0.58	0.88	1.07		0.79	0.57	0.58	0.91		0.77
w	E	0.47	1.32		1.55	0.86	0.46	1.04		1.14	0.78
	I	0.57	3.25		2.86	2.34	0.58	2.98		2.96	2.60
	M	0.47	10.65		11.50	0.82	0.47	0.88		0.90	0.80
	Q	0.58	3.67		3.89	0.80	0.57	2.75		2.94	0.79
x	E	0.47		2.22	9.01	1.14	0.46		2.32	9.44	1.09
	I	0.57		0.90	3.49	2.15	0.58		0.96	4.23	2.87
	M	0.47		0.65	1.81	0.76	0.47		0.61	1.30	0.67
	Q	0.58		1.03	0.97	0.77	0.57		0.88	0.95	0.77
z	E	0.47	1.64	0.71	1.72	1.02	0.46	1.53	0.81	1.54	1.07
	I	0.57	1.73	0.69	1.89	0.70	0.58	1.13	0.64	1.14	1.04
	M	0.47	1.36	0.53	1.31	0.74	0.47	1.05	0.52	1.04	0.71
	Q	0.58	1.89	1.08	1.79	0.72	0.57	1.26	0.88	1.28	0.73

次に、費用総額と他 1 項目が欠測となる a,b,c,d パターンにおいて、売上金額の時点調整及び外れ値処理の有無の違いが補完値に及ぼす影響について比較を行う。このシミュレーション結果を表 9 及び 10 に示す。表 9 は売上金額の時点調整を行わず、売上金額を横置きで補完した後、比率ホットデック補完を行った結果である。表 10 は、売上金額の時点調整を行い、比率ホットデック補完を行った結果である。それぞれ、比率ホットデック補完を行う際のドナーデータの外れ値処理の有無別の結果を表に示す。

表 9　欠測パターン a～d の産業別（売上金額の時点調整なし）

パターン	産業	外れ値処理なし					外れ値処理あり				
		費用総額	棚卸期首	仕入金額	棚卸期末	経費計	費用総額	棚卸期首	仕入金額	棚卸期末	経費計
a	E	0.11	0.85				0.12	0.97			
	I	1.10	7.22				0.42	3.16			
	M	0.25	5.42				0.16	2.27			
	Q	0.13	2.31				0.18	1.94			
b	E	0.46		0.81			0.86		1.32		
	I	0.43		0.48			0.70		0.86		
	M	0.33		0.69			0.36		0.77		
	Q	0.67		1.34			0.72		1.59		
c	E	0.46			0.33		0.86			0.48	
	I	0.43			0.98		0.70			1.03	
	M	0.33			1.15		0.36			1.44	
	Q	0.67			1.12		0.72			1.94	
d	E	0.71				1.21	0.71				1.32
	I	0.19				1.11	0.18				0.84
	M	0.51				0.87	0.62				1.06
	Q	0.74				1.18	0.79				1.17

表 10　欠測パターン a～d の産業別（売上金額の時点調整済み）

パターン	産業	外れ値処理なし					外れ値処理あり				
		費用総額	棚卸期首	仕入金額	棚卸期末	経費計	費用総額	棚卸期首	仕入金額	棚卸期末	経費計
a	E	0.12	1.06				0.12	1.08			
	I	0.10	1.43				0.45	3.43			
	M	0.15	2.28				0.13	2.03			
	Q	0.04	0.94				0.17	2.13			
b	E	0.47		0.82			0.85		1.29		
	I	0.60		0.72			0.87		1.07		
	M	0.27		0.53			0.34		0.71		
	Q	0.54		1.14			0.50		1.18		
c	E	0.47			0.50		0.85			0.57	
	I	0.60			0.64		0.87			1.18	
	M	0.27			1.25		0.34			1.50	
	Q	0.54			1.88		0.50			1.82	
d	E	0.68				1.23	0.72				1.36
	I	0.21				0.92	0.18				0.84
	M	0.54				0.96	0.61				1.06
	Q	0.62				0.91	0.81				1.18

表 9 では、外れ値処理の有無別で真値と補完値との差異は大きく変わるところは見られなかった。ただし、期首棚卸高及び期末棚卸高は産業により真値と補完値の差異が大きくなるものもある。売上金額の時点調整を行った表 10 でも表 9 と同様、外れ値処理の有無別の比較では、大きな差はみられなかった。売上金額の時点調整の有無については、表 9 と 10 を比較すると、項目により誤差が大きくなるものと小さくものがあり、傾向としてはどちらも同じくらいの誤差のように見られた。

使用したドナーレコードの外れ値処理前と処理後の分布を、産業別に図 5～12 に示す。全ドナーレコードの分布において外れている値が、外れ値処理後では外れ値として検出されていることが見える。

なお、外れ値検出対象項目は、売上金額、期首棚卸高、仕入高、期末棚卸高及び経費計の５変数である。ただし、期首棚卸高及び期末棚卸高の相関が非常に高いために分散共分散行列が特異行列となる場合には、期末棚卸高を除いた４変数で外れ値処理を行っている。

図 5　産業 E 欠測率 20%ドナーレコード

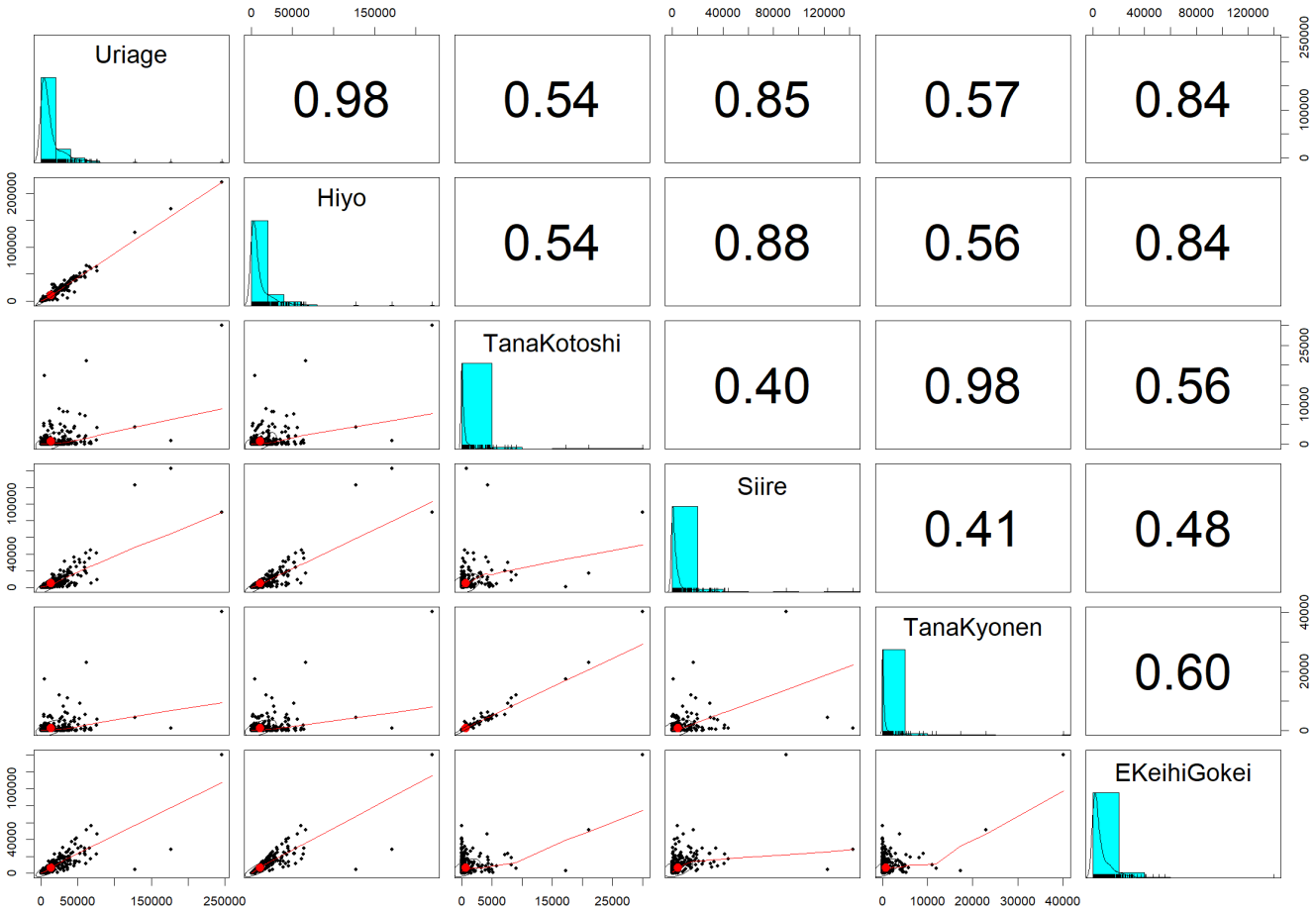


図 6　産業 E 欠測率 20%ドナーレコード　外れ値処理後

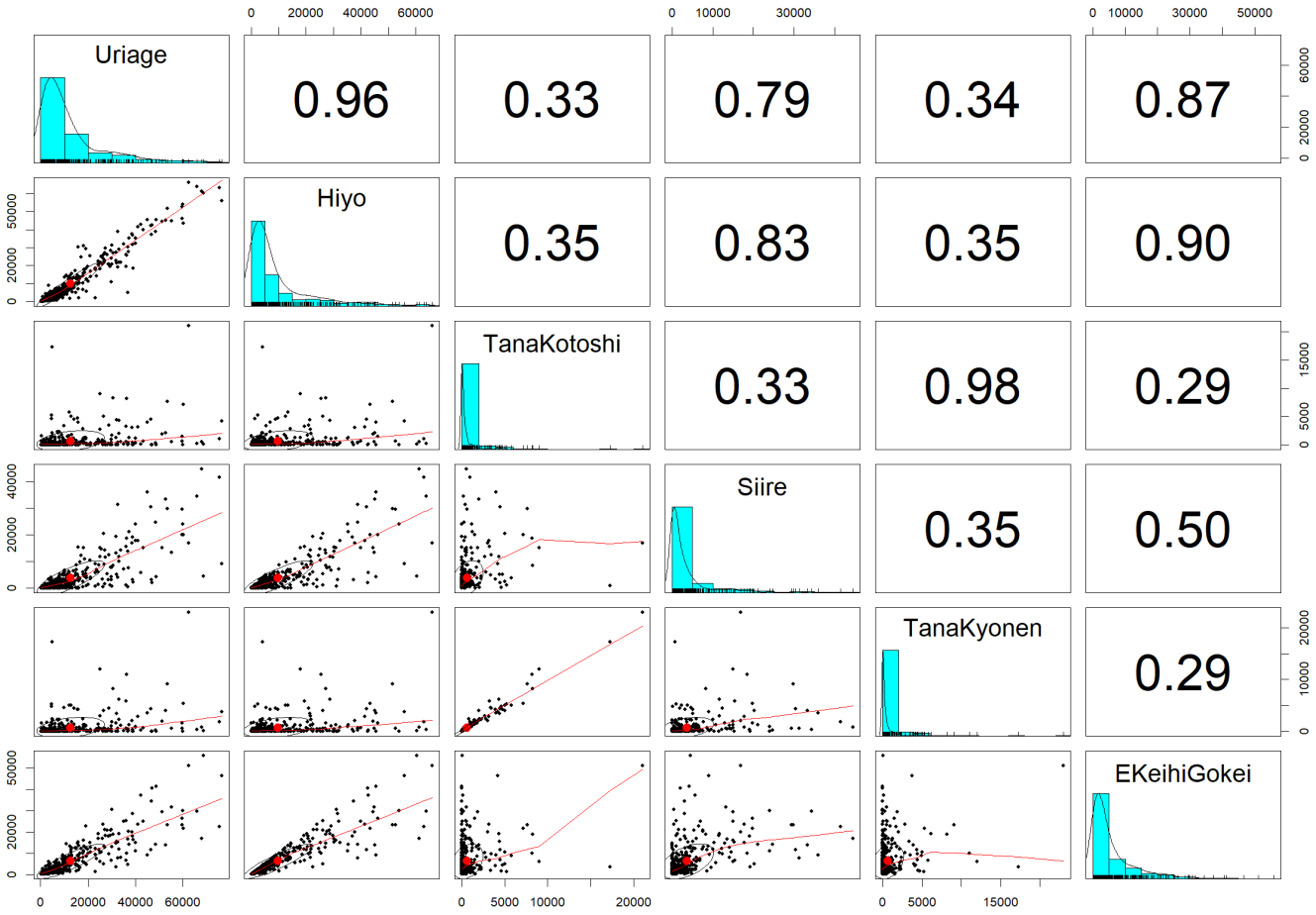


図７　産業Ｉ欠測率 20%ドナーレコード

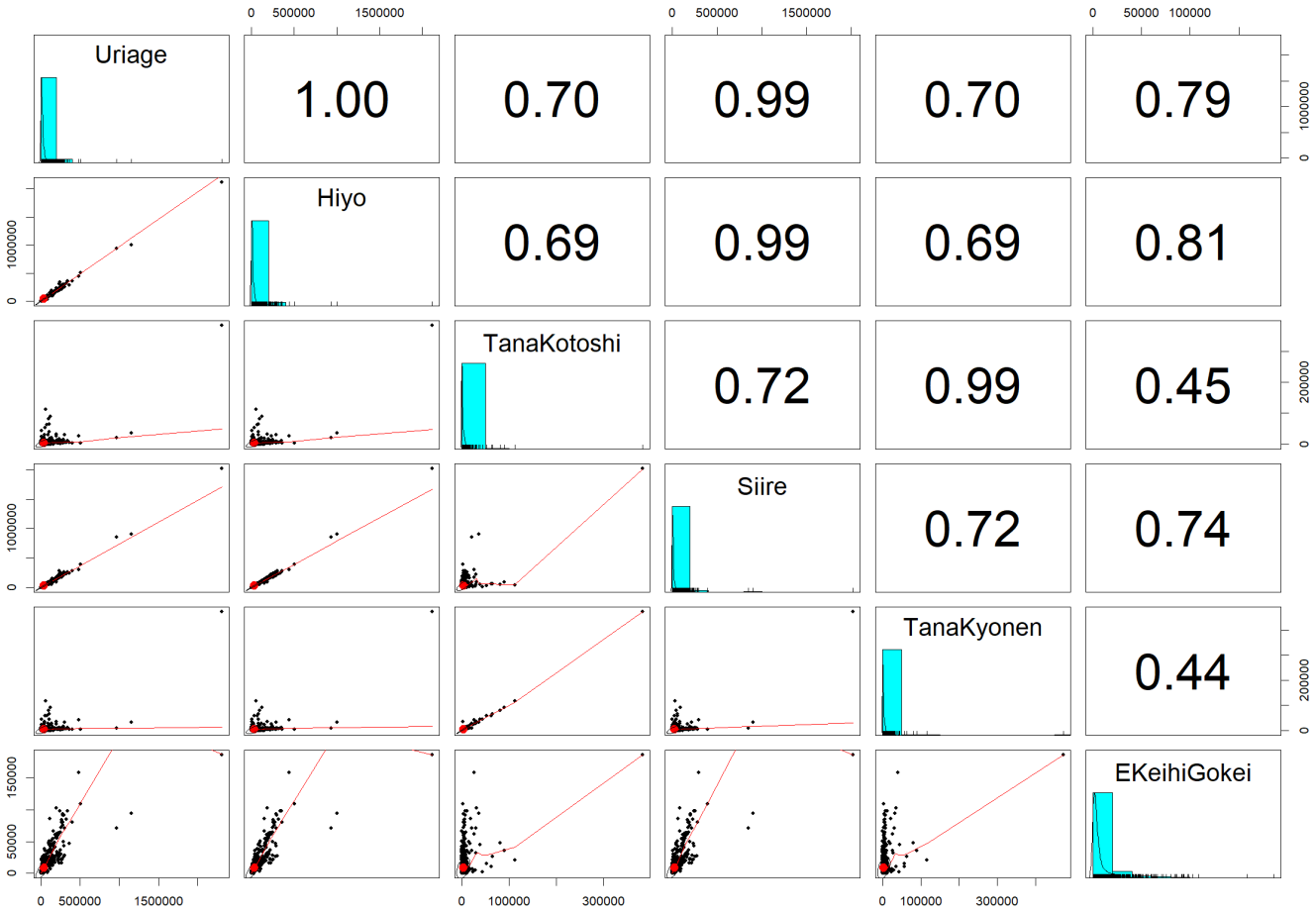


図８　産業Ｉ欠測率 20%ドナーレコード　外れ値処理後

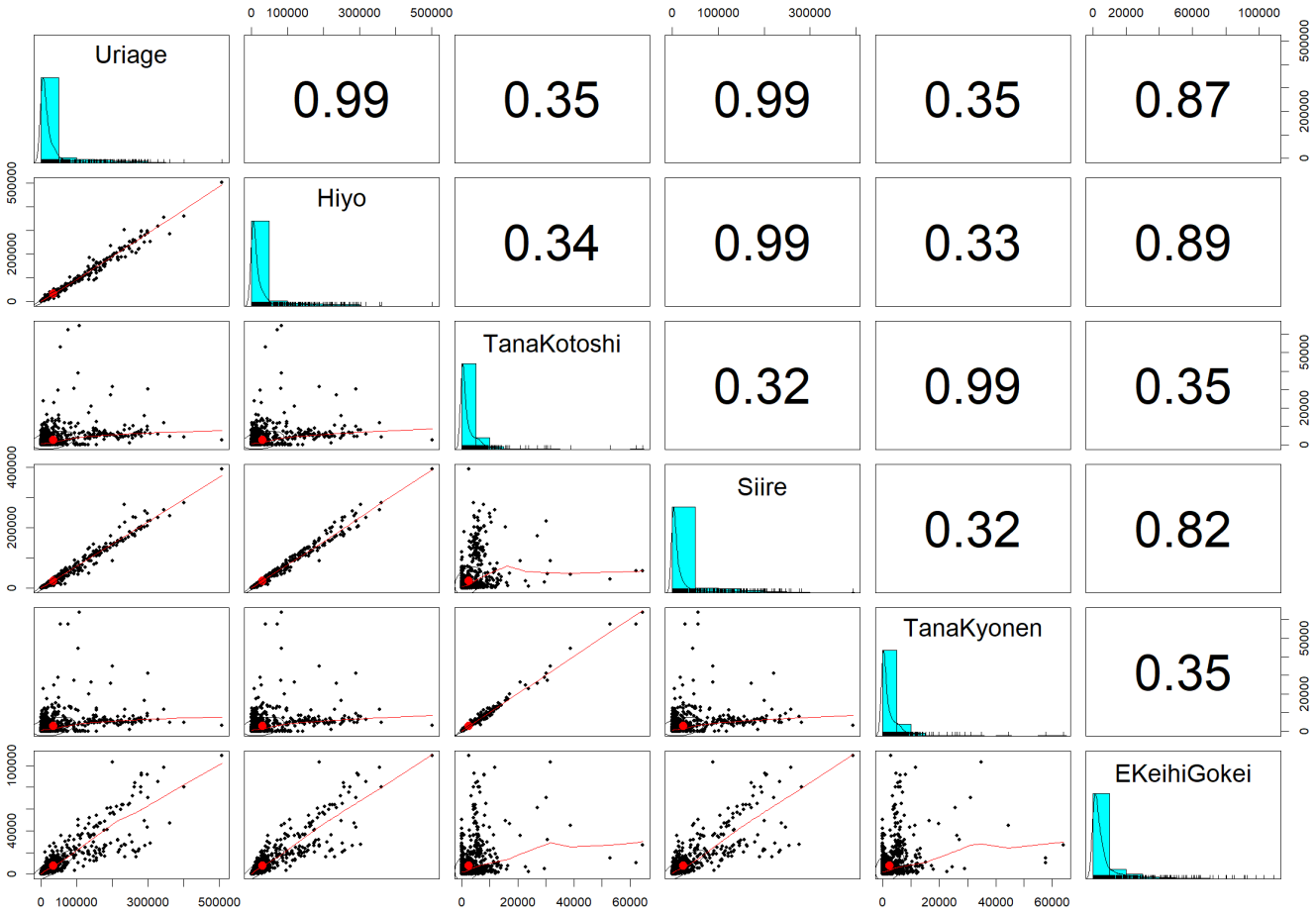


図 9　産業 M 欠測率 20%ドナーレコード

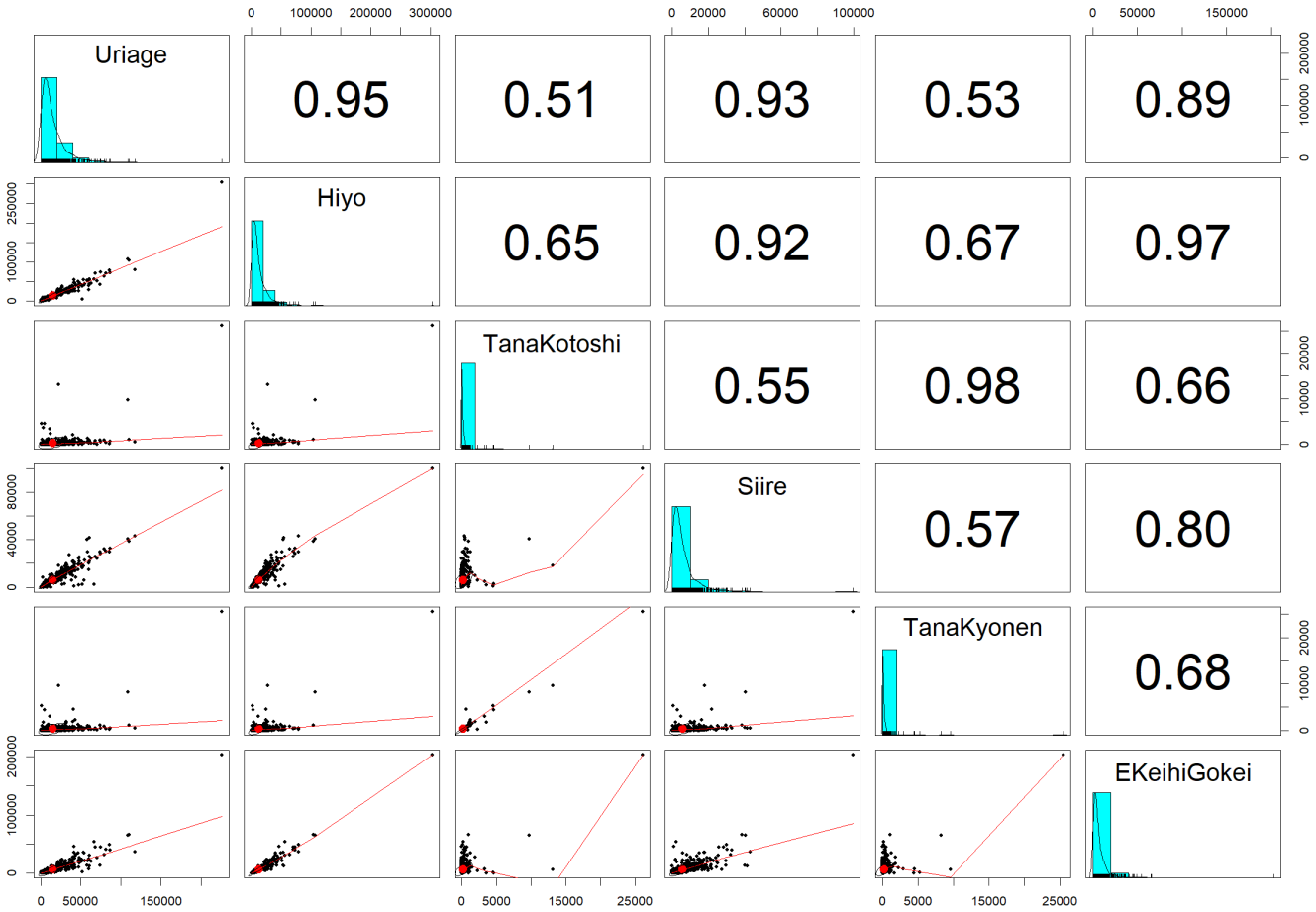


図 10　産業 M 欠測率 20%ドナーレコード　外れ値処理後

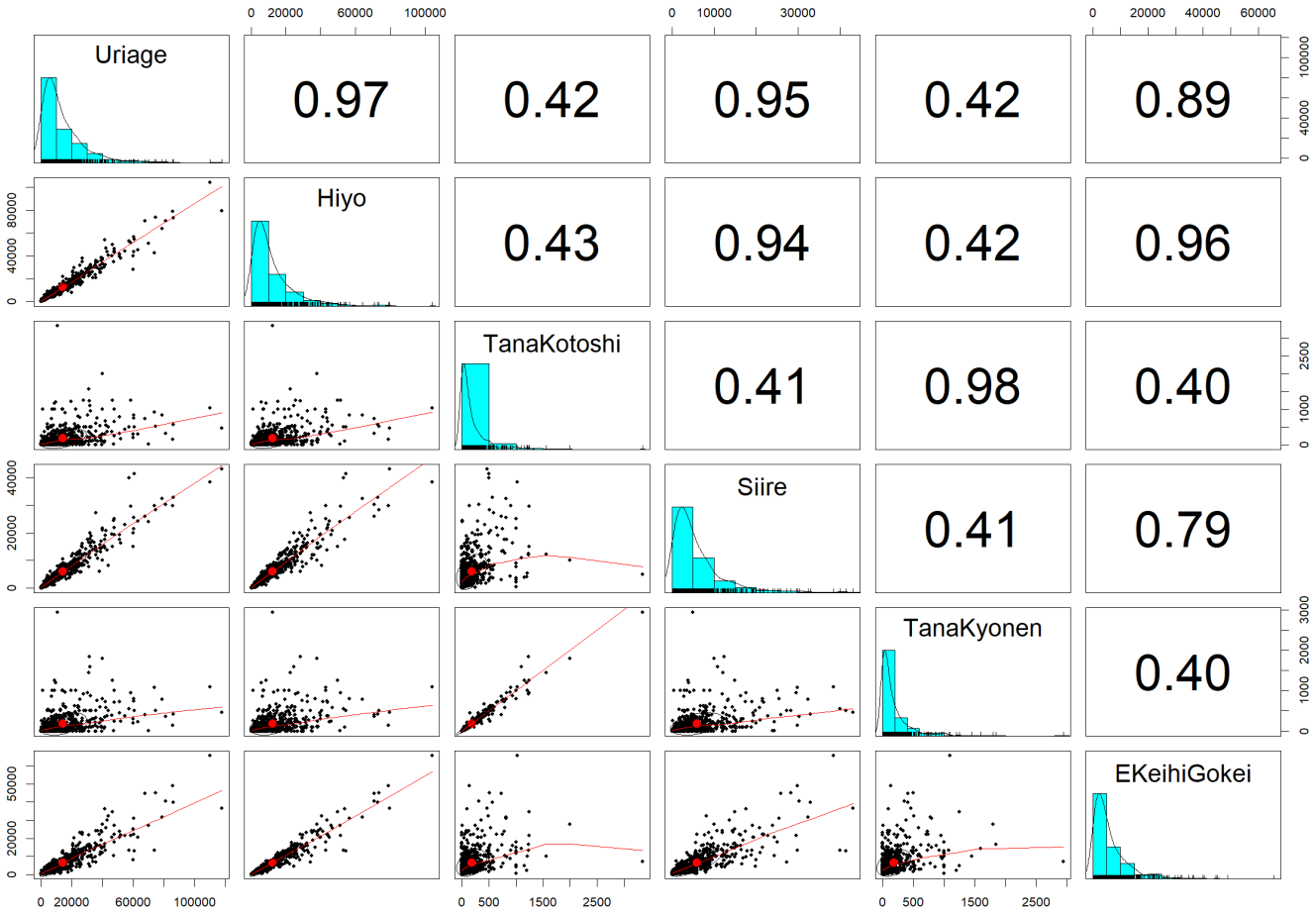


図 11　産業 Q 欠測率 20%ドナーレコード

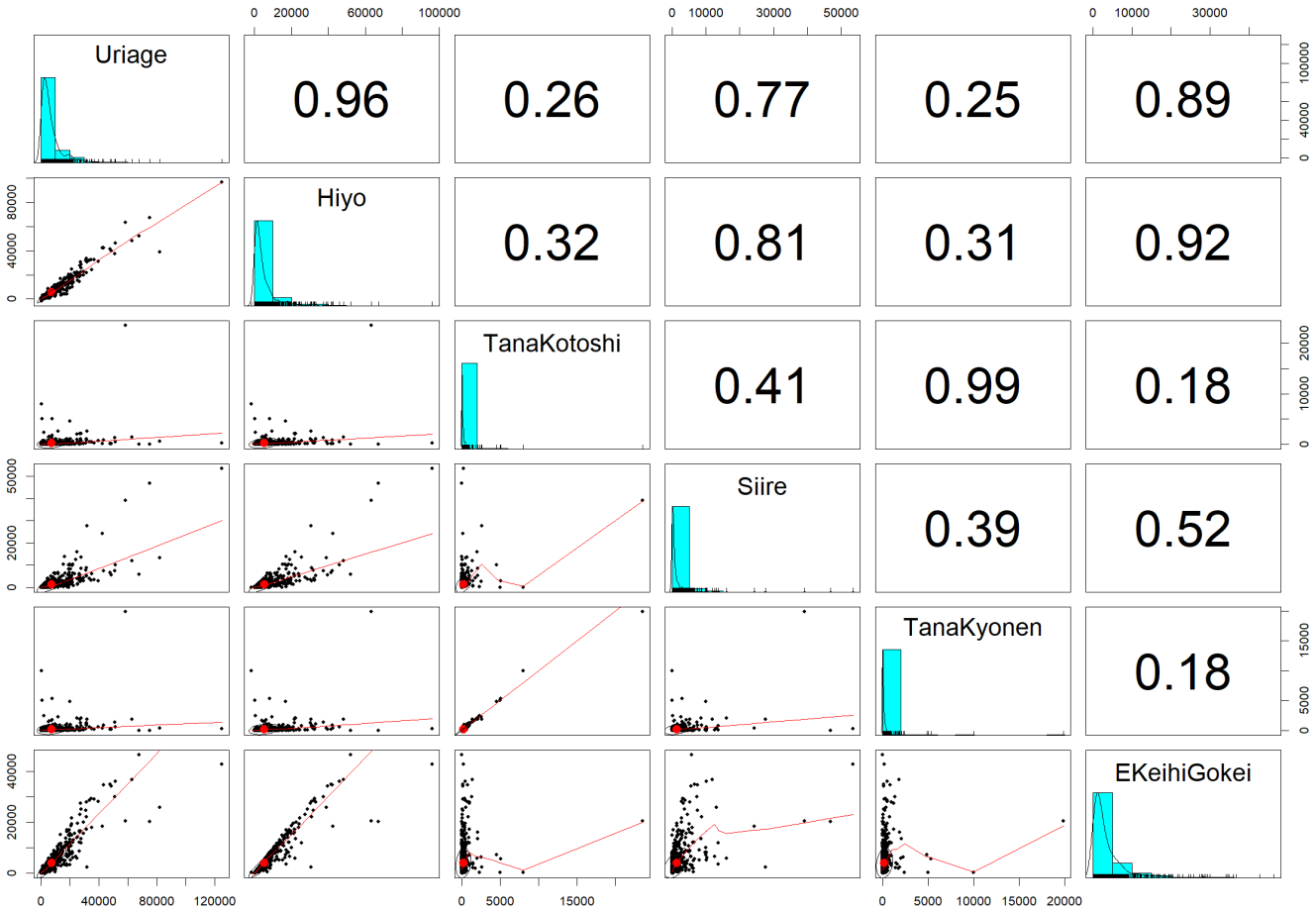


図 12　産業 Q 欠測率 20%ドナーレコード　外れ値処理後