

データサイエンス作法

データサイエンスコンソーシアム, 慶應義塾大学

柴田 里程

データサイエンス

- バズワード
 - ビジネス上の必要に応じてデータを利用するパラダイム(Davenport & Patil)
 - データ・サイエンティスト
- 科学
 - データの科学
 - 理論
 - 実践
 - ソフトウェア
 - 統計科学

データサイエンス作法

- 失敗しないための心得
 - お茶の作法, 食事の作法 (manners, 所作)
- サイエンスの反映
- データの活用にあたって押さえるべきポイント
 - 機械学習
 - 何をどんな風に学習させたいのか
 - どんな入出力データが望ましいのか
 - XAI (Explainable AI)
 - オープンデータ
 - 何をどう公開したらよいのか

データの利用と活用

- 利用

- 都合のよいデータをピックアップするだけ
 - すでに結論は出ているが、お飾りとしてデータが必要
- 特にサイエンスは必要としない
 - 本当に目的に合ったデータかどうか疑問を持たなければ

- 活用

- データを活かし切る, 本当の価値を引き出す
 - データの全体像を把握する
 - 可能な活用形態を調べ上げる
 - より適切なデータ取得

資料, 情報, データ

- 資料
 - 何らかの記録の集まり
 - 存在することに意味がある
- 情報
 - 何らかの目的をもってまとめられた資料
 - 情報ベース
- データ
 - 記号で表された, 推論の根拠となる事実の反映
 - 論拠, 論拠の前提条件

『データサイエンスの作法 ―データの活用を助ける科学のツボ―』

近代科学社, 2020

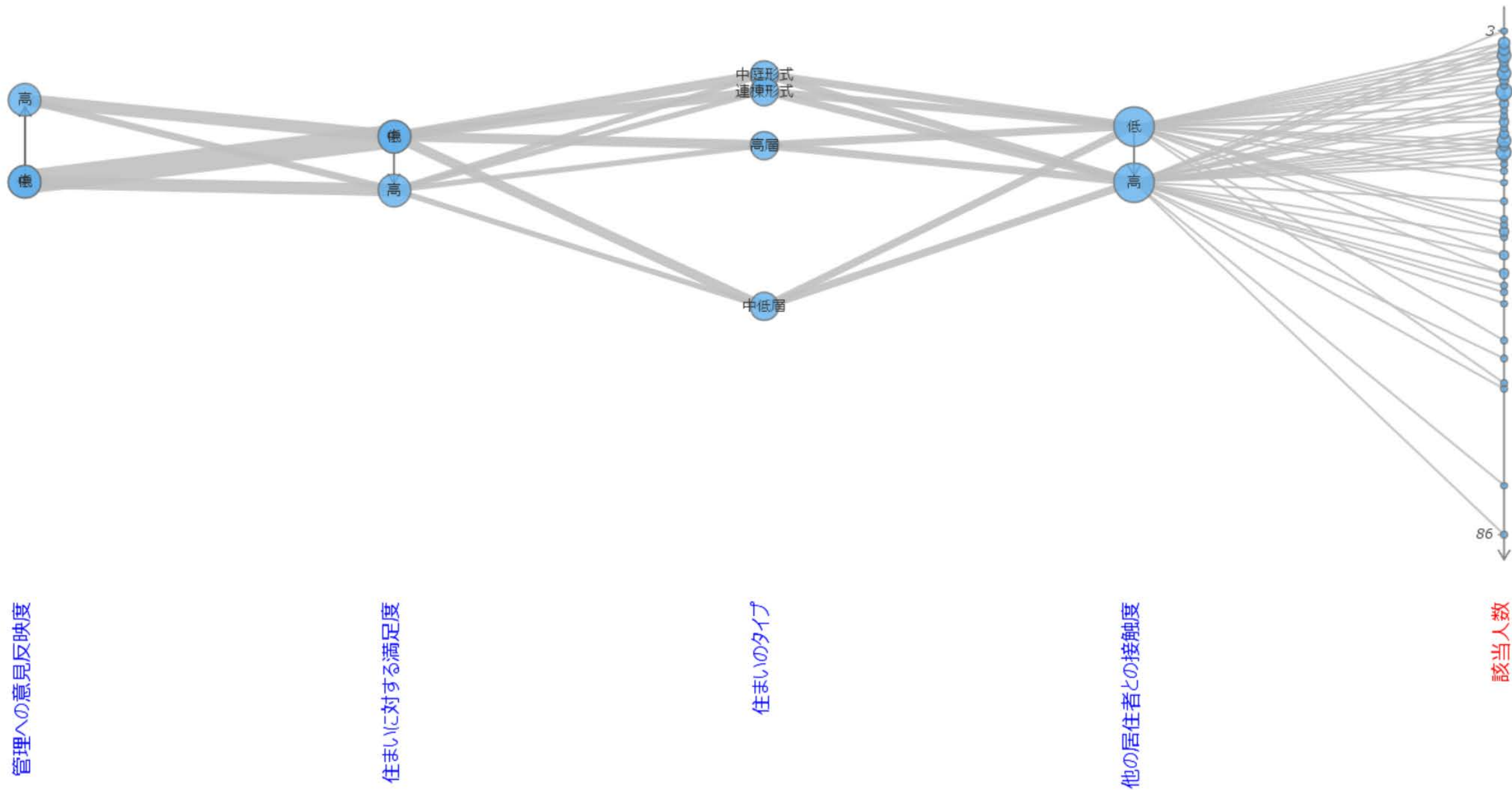
- 視覚表示
- フィルタリング
 - 変量フィルタリング
 - 記録フィルタリング
- 型
- 読込み
- 射影
- 変容
 - 正規化
 - 結合
 - 分解
- R とその利用

記録度数型

4次元分割表

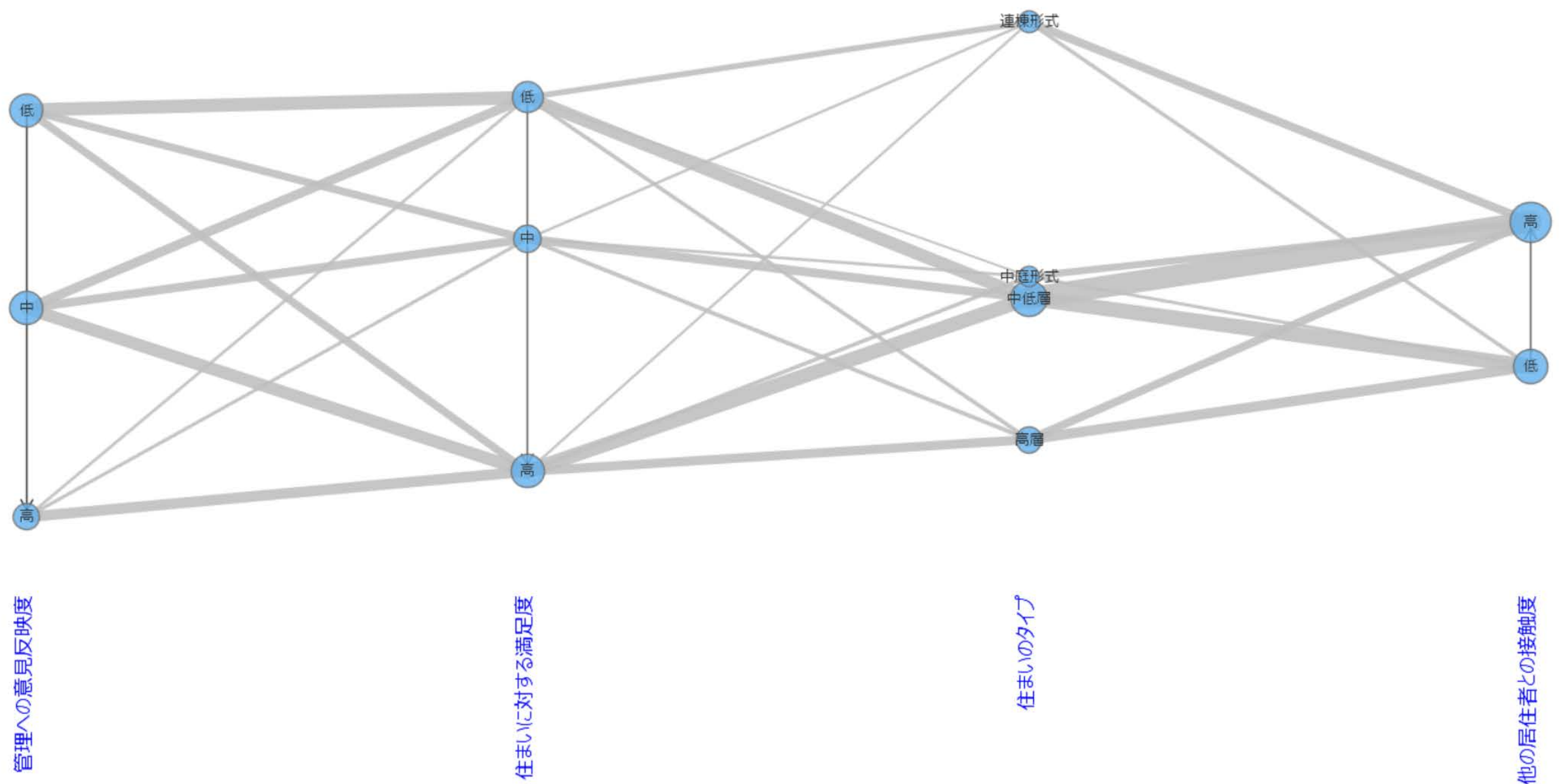
賃貸集合住宅の満足度

居住者



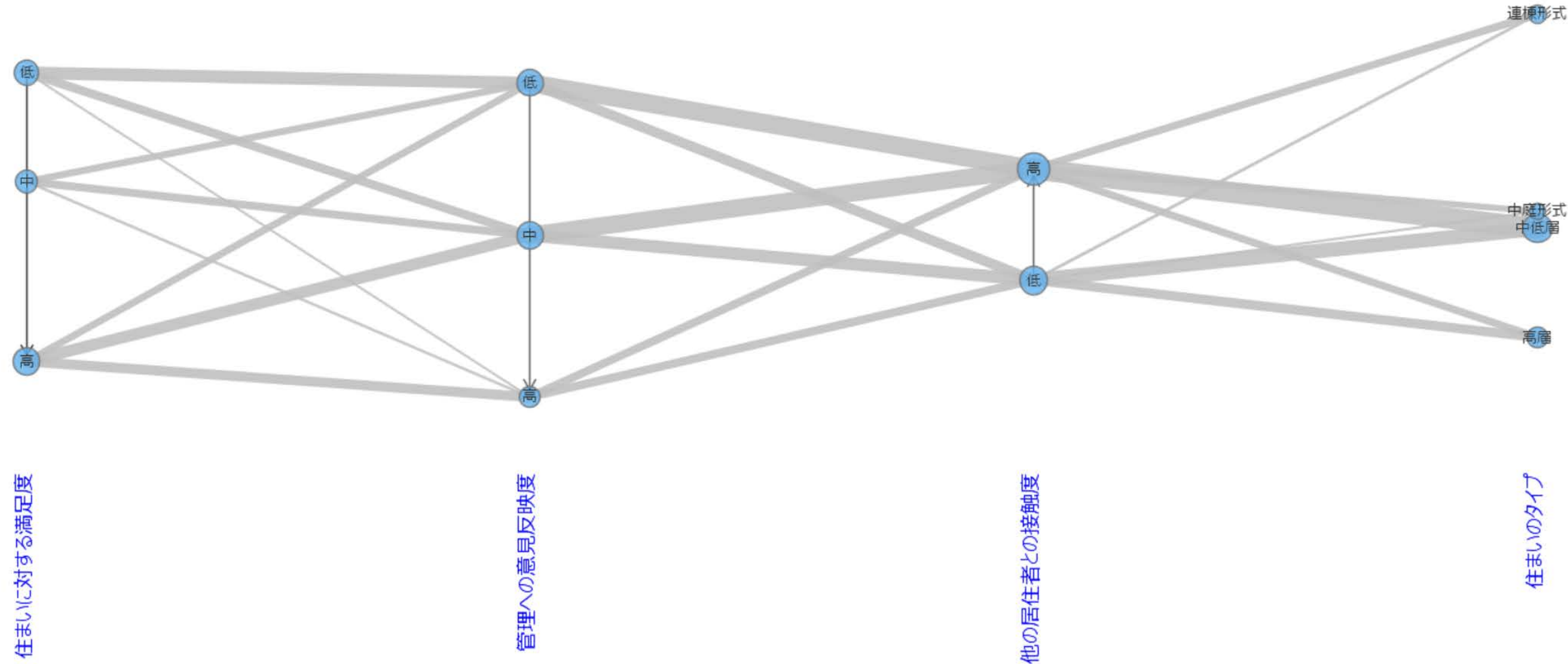
賃貸集合住宅の満足度.1

住人



賃貸集合住宅の満足度.1

住人



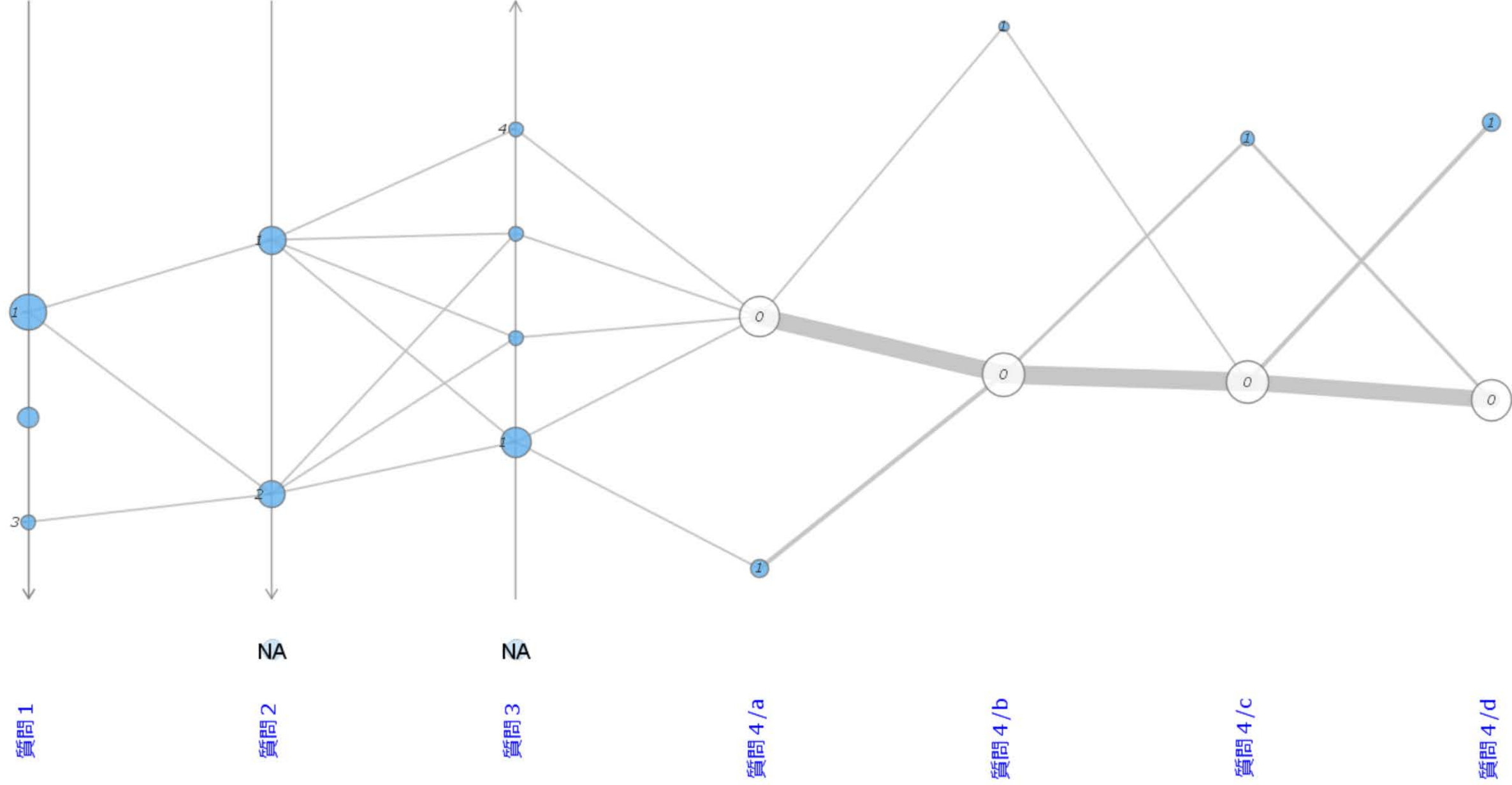
正規化

隠れた変量

アンケートデータ(demo)

質問1	質問2	質問3	質問4/a	質問4/b	質問4/c	質問4/d
1	2	1	0	0	0	0
1	1	4	0	0	1	0
1	1	1	0	0	0	0
2			0	0	0	0
1	1	3	0	0	0	1
1	1	1	0	0	0	0
1	1	1	0	0	0	1
1	2	3	0	0	0	0
1	2	1	1	0	0	0
3	2	1	1	0	0	0
1	1	2	0	0	1	0
1	2	1	0	0	0	0
2			0	1	0	1
2			0	0	0	0
1	1	4	0	0	0	0
3	2	1	0	0	0	0
1	2	2	0	0	0	0
1	1	1	1	0	0	0
2			0	0	0	0

demo



データベクトルの属性などの編集

Id	dv1	dv2	dv3	dv4	dv5	dv6	dv7
短名	v1	v2	v3	v4	v5	v6	v7
長名	質問 1	質問 2	質問 3	質問 4 /a	質問 4 /b	質問 4 /c	質問 4 /d
単位							
型	計測値	計測値	計測値	論理	論理	論理	論理
マーク値				2	2	2	2
説明							
穴埋め							

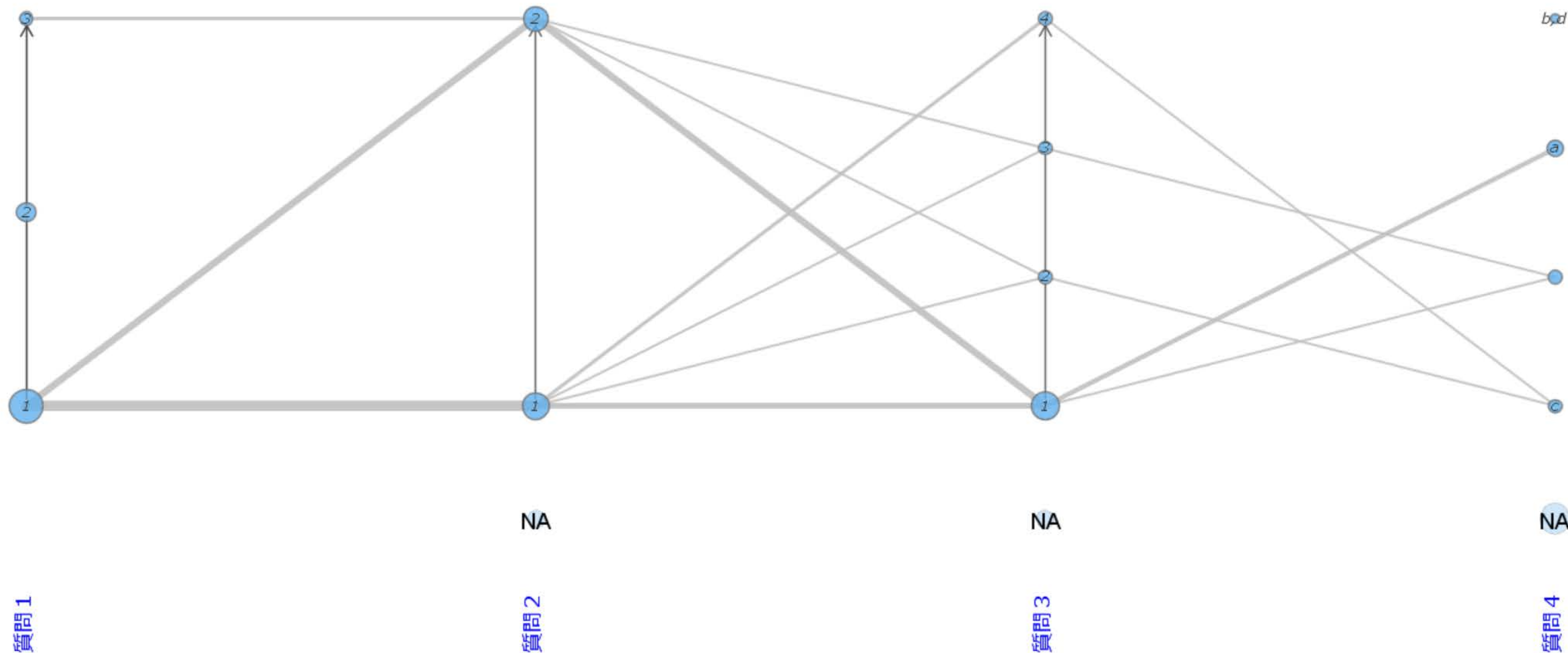
OK



データベクトルの属性などの編集

Id	dv1_1	dv2_1	dv3_1	dv4	dv5	dv6	dv7
短名	v1	v2	v3	v4	v5	v6	v7
長名	質問 1	質問 2	質問 3	質問 4 /a	質問 4 /b	質問 4 /c	質問 4 /d
単位							
型	順マーク	順マーク	順マーク	論理	論理	論理	論理
マーク値	3	2	4	2	2	2	2
説明							
穴埋め							

OK

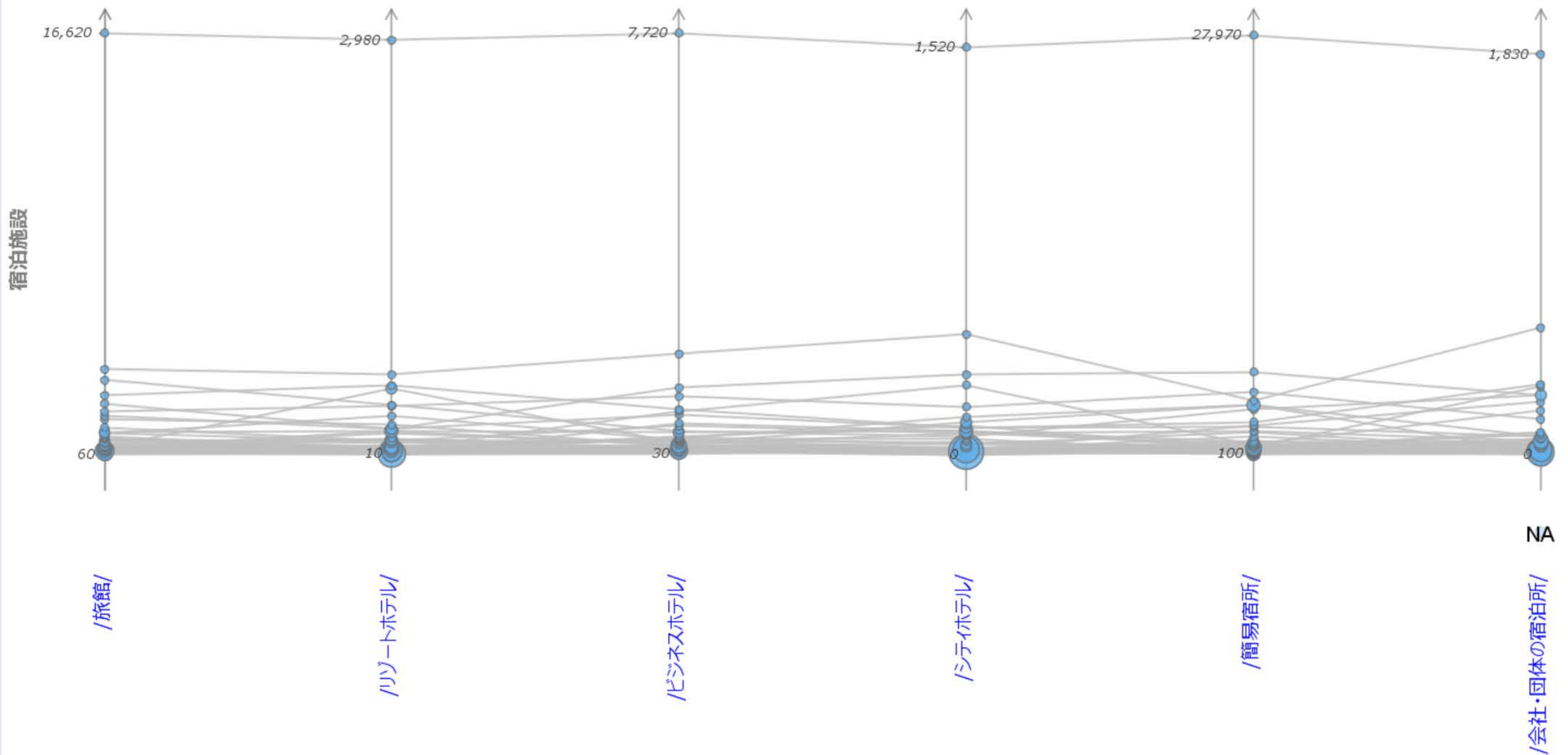


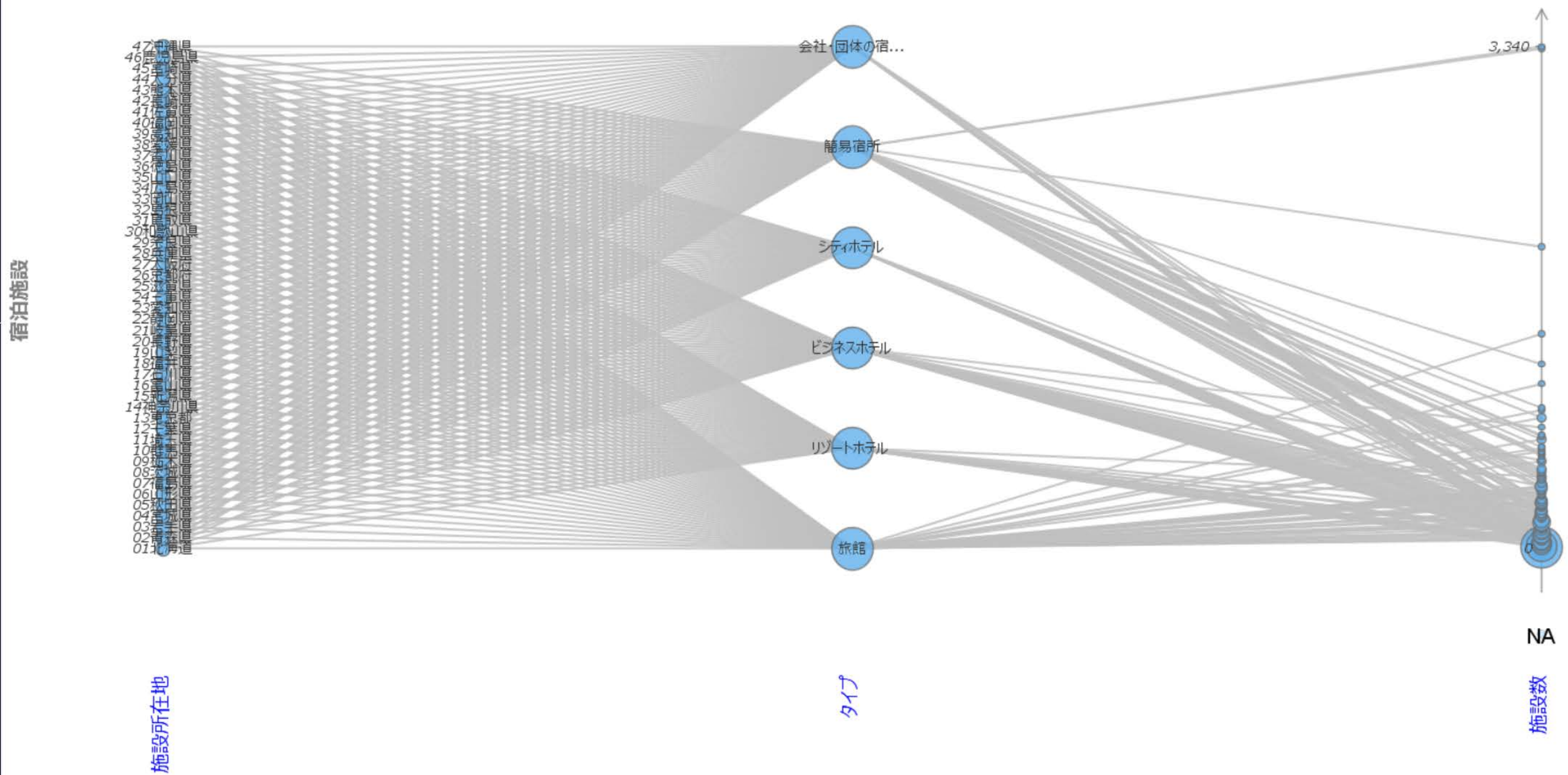
宿泊旅行統計調査第1表

第 1 表 施設所在地(47区分及び運輸局等)、従業者数(4区分)、宿泊目的割合(2区分)別施設数
並びに施設所在地(47区分及び運輸局等)、宿泊施設タイプ(6区分)別施設数

施設所在地(47区分 及び運輸局等)	従業者数(4区分)、宿泊目的割合(2区分)															(施設)					
	総数 1)、2)	宿泊目的割合(2区分)														宿泊施設タイプ(6区分)					
		観光目的の 宿泊者が 50%以上	観光目的の 宿泊者が 50%未満	0～9人			10～29人			30～99人			100人以上			旅館	リゾート ホテル	ビジネス ホテル	シティ ホテル	簡易宿所	会社・団体の 宿泊所
				1)	観光目的の 宿泊者が 50%以上	観光目的の 宿泊者が 50%未満	1)	観光目的の 宿泊者が 50%以上	観光目的の 宿泊者が 50%未満	1)	観光目的の 宿泊者が 50%以上	観光目的の 宿泊者が 50%未満	1)	観光目的の 宿泊者が 50%以上	観光目的の 宿泊者が 50%未満						
施設所在地 計	58,989	41,860	13,470	45,360	34,070	7,770	7,950	4,270	3,640	4,496	2,790	1,700	1,094	730	360	16,620	2,980	7,720	1,520	27,970	1,830
01北海道	2,556	1,360	1,060	1,850	980	730	378	160	210	248	160	90	76	60	20	920	170	420	90	870	60
02青森県	527	240	280	362	160	190	102	40	60	53	30	30	9	10	0	260	10	90	20	130	10
03岩手県	682	390	280	471	300	160	118	40	80	82	40	40	9	10	0	290	30	80	30	250	0
04宮城県	689	390	280	446	280	150	145	60	80	76	30	40	21	20	0	230	20	180	20	230	10
05秋田県	558	340	190	380	240	110	116	70	50	55	30	20	7	0	0	230	10	50	10	250	0
06山形県	868	610	220	646	470	140	142	90	50	69	40	30	11	10	10	380	40	90	20	320	20
07福島県	1,787	1,040	570	1,417	830	410	240	120	120	104	70	40	25	20	10	650	90	260	10	730	30
08茨城県	627	290	320	429	230	190	123	30	90	61	30	30	12	0	10	220	10	190	10	180	20
09栃木県	1,252	950	210	939	750	100	179	120	60	116	70	40	15	10	10	390	100	120	10	530	100
10群馬県	1,003	750	220	757	580	140	147	100	50	86	60	30	13	10	-	670	30	80	10	190	10
11埼玉県	353	160	180	196	130	50	82	10	70	63	20	50	12	0	10	70	10	130	20	130	0

タイプ別施設数.2020.02





タイプ

- ビジネスホテル
- 旅館
- 会社・団体の宿泊所
- シティホテル
- リゾートホテル
- 簡易宿所

