テキストマイニングを用いた自由記述データの有効活用に関する研究
―群馬県南牧村を対象として―

群馬工業高等専門学校　環境都市工学科　学生会員　諸岡 峻一
群馬工業高等専門学校　環境都市工学科　正会員　森田 哲夫
前橋市都市計画部まちづくり課　正会員　塚田 伸也

１．研究の背景・目的

(a)研究の背景

都市計画などにおいて、住民の地域に対するイメージを把握することは重要である。従来から地域のイメージに関する調査・研究が進められていて計画分野の研究者や自治体の重要な課題の１つとなっている。また、アンケート調査では、自由記述欄が設けられている場合が多い。最近ではテキストマイニングなどにより自由記述アンケートなどを定量的に分析できるようになってきている。

(b)研究の目的

本研究では、テキストマイニングを用いてアンケート調査の自由記述データを分析し、その結果をプリコードデータに補充することで住民による生活質評価と居住意向の関係を明らかにする。これにより、自治体などのアンケート調査における自由記述データの有効活用の方法を提案することを目的とする。

２．本研究の位置づけ

(a)既存研究の整理

テキストマイニングを用いて自由記述データを分析した研究として、森田・入澤ら[1)]の研究があり、群馬県前橋市の居住者に対するアンケート調査を行い、自由記述データとプリコードデータの関係を分析している。限界自治体を対象地域とした森田・下風ら[4)]の研究では群馬県南牧村を対象に生活質アンケートを実施し、共分散構造分析を用いて、生活質と居住意向の関係を分析している。

キーワード　自由記述アンケート、テキストマイニング、限界自治体
連絡先　〒371-8530　群馬県鳥羽町 580
群馬工業高等専門学校環境都市工学科
TEL:027-254-9179
E-mail：tmorita@cvl.gunma-ct.ac.jp

(b)本研究の位置づけ

森田・下風ら[4)]の研究では、生活質と居住意向の関係を分析するために共分散構造分析を行い、居住意向を決定するものとして年齢や居住歴の他に被災歴、職業、主要施設への道のり等が影響していることが分かっているが、生活質の総合評価との関係は示されなかった。本研究では、このモデルに、自由記述データから得られる因子を加えることで、生活質の総合評価と居住意向の関係を明らかにしていく。これにより、アンケート調査の自由記述データを政策評価などに活用できるのではないかと考える。

本研究の分析フローを図 1 に示す。自由記述データの頻出語の出現数の集計、クラスター分析、共起ネットワーク作成には、テキストマイニングのフリーソフトウェアである KH coder を使用している。

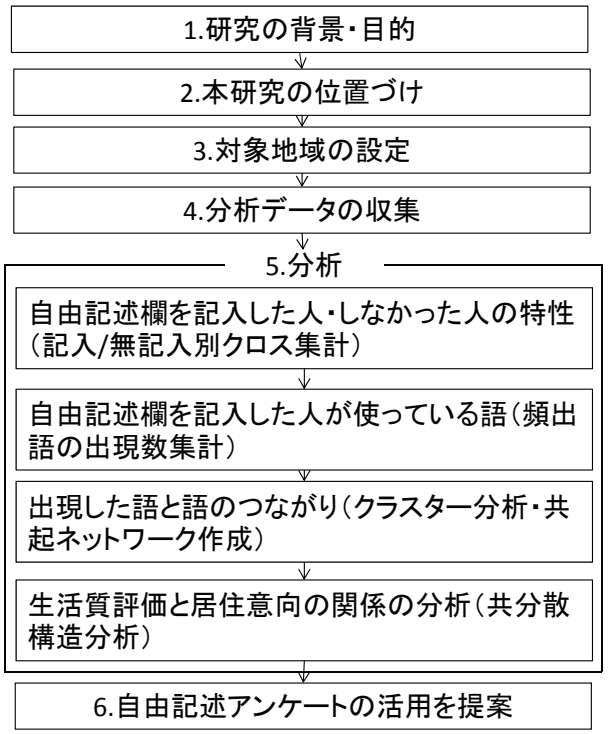

図 1　研究のフロー

３．対象地域の設定

本研究では、群馬県南牧村を対象地域とする。南牧村は限界自治体の中でも最も高齢化率が高く、早急に生活質を維持するための計画、あるいは計画的な集落撤退を考えなくてはならない地域である。南牧村の全図を図 2 に示す。

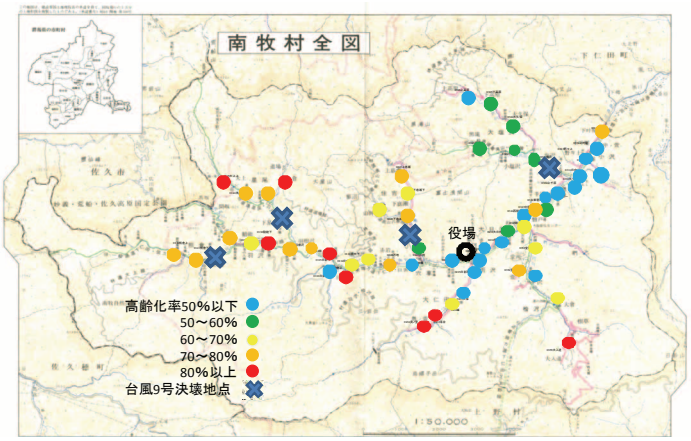

図 2　南牧村全図

４．分析データの収集

本研究では、2010 年に群馬県南牧村で行われたアンケート調査を用いる。アンケート調査の企画概要を表 1 に示す。

表 1　企画概要

調査日	配布:2010年11月1日
	回収:2010年11月21日(郵送投函期限)
対象地域	群馬県甘楽郡南牧村全域
対象者	全1,117戸の世帯主あるいは代表者
調査方法	配布:集落代表者による戸別配布
	回収:郵送回収
調査内容	1)個人属性(性別、年齢、職業、運転免許の有無)
	2)世帯属性(世帯構成、住宅形式、所持自動車数)
	3)災害による被災経験(2007年9月台風9号)
	4)生活の質評価(23項目、総合評価):五段階評価
	5)居住意向(居住年数、定住/転居意向、転居理由)
	6)自由記述(南牧村のイメージ)
回収数	配布数:1,117標、回収数:637票、回収率:57.0%
	うち有効票483票
調査主体	群馬工業高等専門学校　環境都市工学科
	群馬県　県土整備部　都市計画課

５．分析

(a) 自由記述欄を記入した人・しなかった人の特性

アンケート調査の自由記述欄を記入した人・しなかった人がそれぞれどのような人であるかを把握するために自由記述欄の記入・無記入別に世帯属性、被災経験、居住意向、地区特性の各項目でクロス集計を行った。図 3～図 5 から、世帯属性の項目では、記入している人の方が 60 代以下の若い年代の割合が多く、就業者の割合が多いことがわかる。このことから、記入した人は時間があるから記入したというよりは、何か書きたいことがあったと考えられる。就業者の仕事場所は南牧村以外の割合が多かったため、他の市町村と比較して、南牧村が不便だと感じ、記入しようとした可能性も考えられる。また、図 6 から、居住意向の項目では、記入している人の方が現在地に住み続けるという人の割合が少ないことがわかり、現在、居住している場所に不満があって自由記述欄を記入した可能性も考えられる。被災経験、地区特性の項目に関しては記入・無記入で、違いはみられなかった。

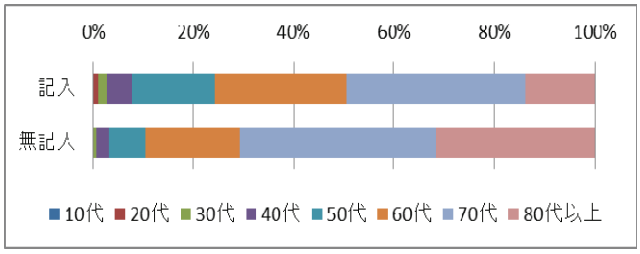

図 3　記入・無記入別の年代

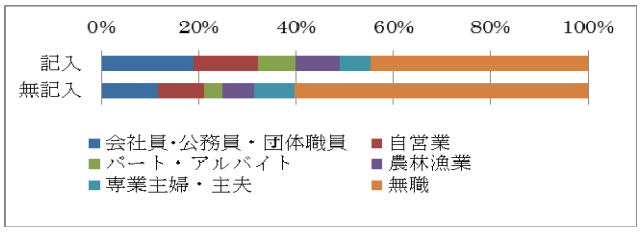

図 4　記入・無記入別の職業

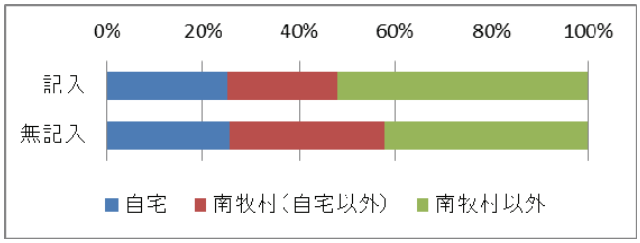

図 5　記入・無記入別の仕事場所

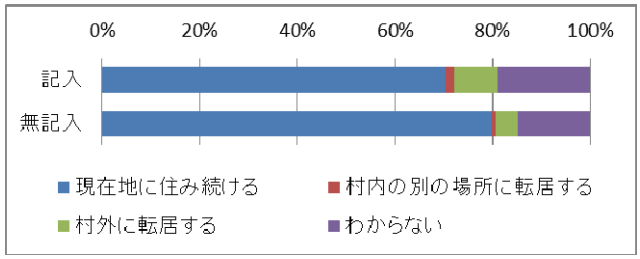

図 6　記入・無記入別の居住意向

(b) 自由記述欄を記入した人が使っている語

記入している人が、どのような語を使っているかを把握するため、形態素解析を行い、自由記述データ全体の頻出 50 語の出現数を集計した。さらに、プリコードデータの世帯属性、被災経験、居住意向、地区特性の各項目で頻出語を集計した。表 3 に自由記述データ全体の頻出 50 語を示す。色がついている単語は名詞である。“水”“空気”などの自然に関する語や“不便”“高齢化”など限界自治体に関連するような語が多く出現していることがわかる。

表 2　頻出 50 語の出現数

順位	抽出語	出現数	順位	抽出語	出現数
1	村	104	27	緑	18
2	する	67	28	人口	17
3	自然	64	29	高齢者	16
4	ない	63	30	美しい	16
5	多い	52	31	よい	15
6	ない	50	31	川	15
7	水	47	34	場所	14
7	良い	47	34	道路	14
9	思う	42	36	働く	13
9	住む	42	37	おいしい	12
11	空気	40	37	悪い	12
12	人	39	37	恵まれる	12
13	なる	35	37	出る	12
14	きれい	31	41	できる	11
14	山	31	42	環境	10
16	ある	28	42	出来る	10
17	少ない	25	42	心	10
18	生活	24	42	有る	10
19	若い	23	42	老人	10
19	豊か	23	47	近所	9
21	高齢化	22	47	心配	9
21	不便	22	47	進む	9
23	いる	21	47	人達	9
23	子供	21	47	静か	9
25	南牧村	20	47	日本一	9
26	無い	19	47	不安	9
27	過疎	18	47	役場	9

(c) 出現した語と語のつながり

(b)では出現した語と語のつながりを把握できないため、共起ネットワークの作成とクラスター分析を行った。結果をそれぞれ図 7、図 8 に示す。共起の指標は、自由記述データの中で多く出現する中心的な話題を抽出するため、Jaccard 係数を使用した。Jaccard 係数は、語句 X と語句 Y が出現するか否かによって、自由記述データを表 3 のように 4 つの部分集合に分割すると、以下の式で与えられる。

$$Jac(X,Y)=a/(a+b+c)$$

表 3　X と Y の出現に対する分割表

	Yが現れる	Yが現れない
Xが現れる	a	b
Xが現れない	c	d

図 7 から、赤色と水色のクラスターは自然に関する良いイメージの語がつながっていることがわかる。茶色、紫色のクラスターでは高齢化や生活の不便さなどの悪いイメージの語がつながっている。図 8 から、若者の働く場所が少ないこと、道路の状況が悪いことなど、プリコードデータには設定されていない内容も読み取ることができる。

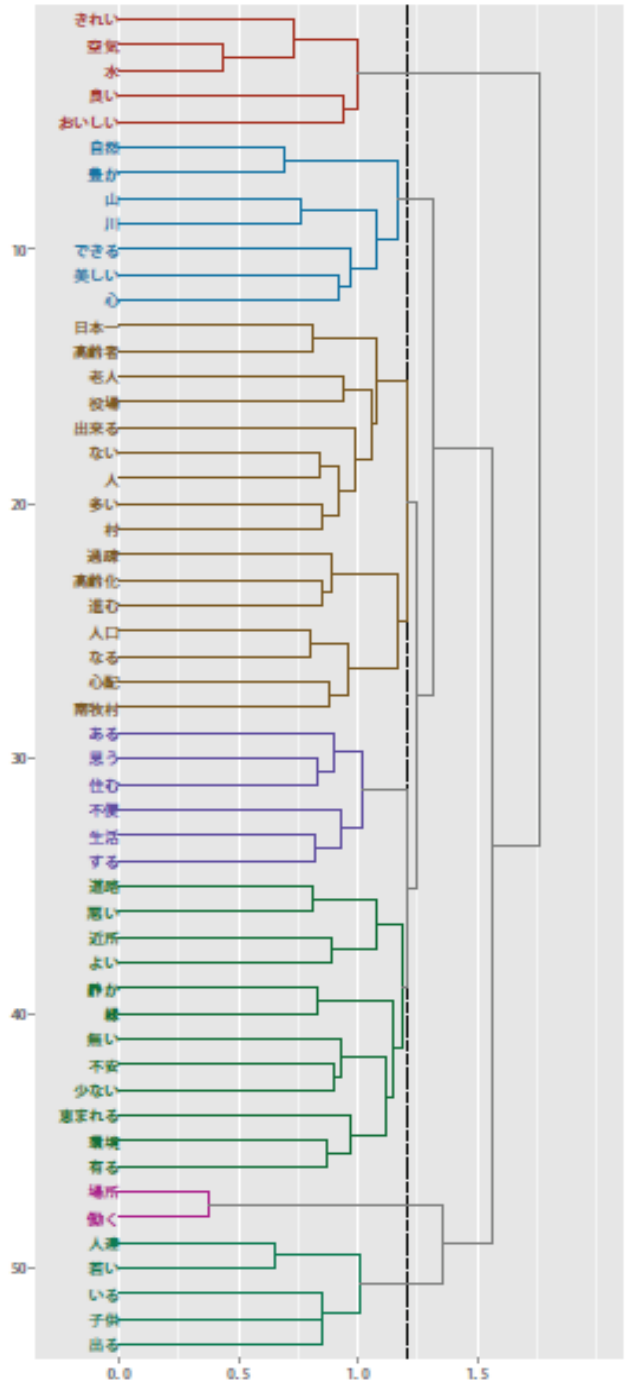

図 7　クラスター分析

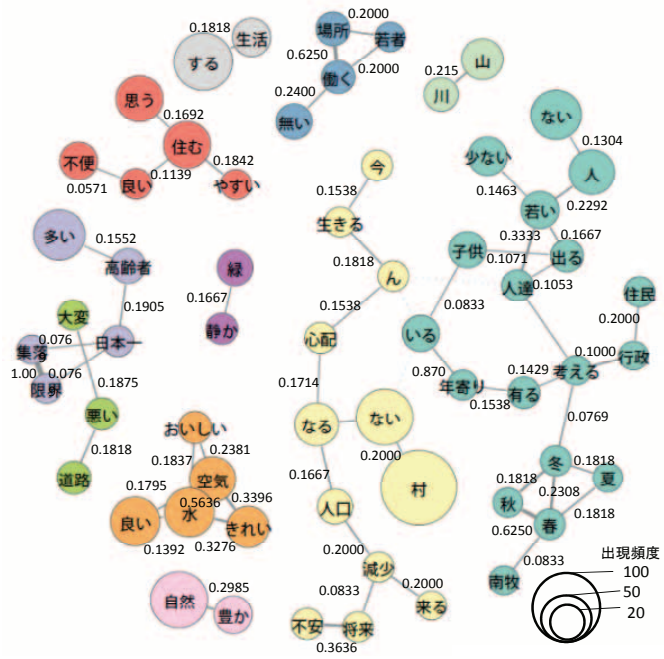

図 8　共起ネットワーク

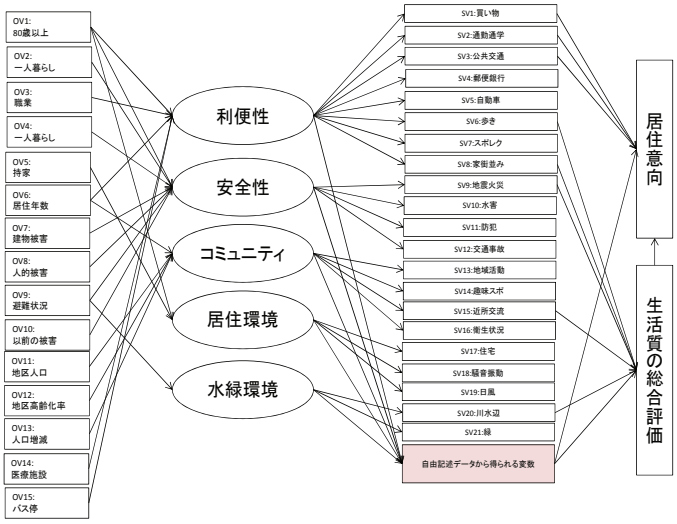

図 9　共分散構造分析パス図

(d) 生活質評価と居住意向の関係の分析

共分散構造分析で自由記述データから得られる変数を加えるために(b)で集計した出現語で、図 7 で分類されたクラスターの色ごとに含まれる語の出現数を集計し、世帯属性、被災経験、居住意向、地区特性の各項目によって出現数に差があるかを調べるため、残差判定を行った。その結果、持家の有無で赤色のクラスター、台風 9 号による人的被害の有無で水色・茶色・紫色・桃色のクラスター、地区人口 60 人以上・60 人未満で茶色のクラスター、地区高齢化率 50%以上・50%未満で水色のクラスターで調整済み残差が 1.96 以上となり有意差があった。この結果を参考に変数を決定し、図 9 のように自由記述データから得られる変数を加えたパス図を作成して共分散構造分析を行う。共分散構造分析の結果については、講演時に詳細を示す。

6. 自由記述アンケートの活用の提案

アンケート調査の自由記述データとプリコードデータの関係を示すことができた。また、自由記述データからは、プリコードデータには設定されていない内容を把握することができた。このことから、プリコードデータを補充する手段として、自由記述データを活用できると考える。

参考文献

1)森田哲夫，入澤覚，長塩彩夏，野村和広，塚田伸也，大塚裕子，杉田浩：自由記述データを用いたテキストマイニングによる都市のイメージ分析，土木学会論文集 D3（土木計画学），Vol.68，No.5（土木計画学研究・論文集第 29 巻），I_315-I_323，2012

2)小林祐司，寺田充伸，佐藤誠治：テキストマイニングを活用したアンケートにおける自由回答の分析と生活環境評価，日本建築学計画系論文集，第 77 巻，第 671 号，pp.85-93，2012

3)佐々木靖弘，佐藤理史，宇津呂武仁：関連用語収集問題とその解法，自然言語処理，Vol.13，No3，pp.150-175，2006

4)森田哲夫，下風笑美子，長塩彩夏：限界自治体における安全性に着目した生活質と居住意向に関する研究，土木学会土木計画学研究・講演集 No.43，CD-ROM（310），2011