第 57 回土木計画学研究発表会・講演集

トピックモデルと離散連続モデルを用いた自由記述の量的分析法

川野　倫輝[1]・佐藤　嘉洋[2]・円山　琢也[3]

[1]学生会員　熊本大学大学院自然科学研究科社会環境工学専攻（〒860-8555 熊本県熊本市中央区黒髪 2-39-1）
E-mail:178d8811@st.kumamoto-u..ac.jp

[2]学生会員　熊本大学大学院自然科学研究科社会環境工学専攻（〒860-8555 熊本県熊本市中央区黒髪 2-39-1）
E-mail:yo-sato@kumamoto-u.ac.jp

[3]正会員　熊本大学准教授　くまもと水循環・減災研究教育センター
（〒860-8555 熊本県熊本市中央区黒髪 2-39-1）
E-mail:takumaru@kumamoto-u.ac.jp

テキストマイニング等の自由記述データの分析手法について，土木計画学分野においても関心が高まっているが分析手法は確立されているとは言い難い．本研究では，トピックモデルと離散連続モデルを統合した新たな自由記述データの分析法を提案する．事例として，2016 年熊本地震による益城町仮設住宅居住者を対象とした聞き取り調査中の自由回答を用いる．まず，トピックモデルにより各自由回答を構成するトピック割合を算出する．次に，このトピック割合をトピックの選択，割合の配分行動と捉え，離散連続モデルを用いた記述を行う．その結果，属性別の回答傾向の違いや，選択式設問の回答と自由回答中のトピック選択の対応が統計的に示され，手法の有用性が確認できた．

Key Words : *free answer, text mining, topic model, MDCEV model*

1. はじめに

近年，機械学習分野で開発されたトピックモデル [1]が注目を集めている．トピックモデルとは，文書データの解析手法であり，教師データなしで文書中からトピックを推定することが可能である．従来から存在するトピック抽出手法と比較し，トピックモデルでは，1）1 つの文書が複数のトピックで構成されることを許容する，2）より多くの語を扱えることから，出現頻度の低い語を含んだ俯瞰的な分析が可能，3）単語の頻度が正の値として算出されるため，自然な解釈が可能といった利点がある．以上の背景から，国内外の土木計画・都市計画分野において，適用事例が見られるようになってきている．学術論文の研究トピックの分析として，Lijun ら [2]は，交通研究分野の 22 の国際誌に 1990 年～2015 年で掲載された 17,163 の論文のアブストラクトを対象にトピックモデルを適用し，研究トピックの増減傾向や研究トピック間の関連を把握している．塚井ら [3]は，1984 年から 2015 年の土木計画学論文集で公開されている 662 の論文にトピックモデルを適用し，研究トピックの時系列変化を把握している．また，会議やワークショップなどの発言データの解析として，塚井ら [4]は，地域公共交通会議の討議録にトピックモデルを適用し，討議中のトピック抽出やトピック推移の把握を通して同手法の適用可能性を検証している．その他，文書データ以外の解析への応用事例もある．神谷ら [5]は，モバイル空間統計データに，文書をメッシュ，滞在者の居住区を単語として，トピックモデルを適用し，地域別人口特性の解釈を行っている．以上のように，土木計画・都市計画分野においてもトピックモデルの適用・応用は見られるものの，モデルから得られるトピック分布を二次的に活用した研究は見当たらない．トピック分布とは，1 つの文書が複数のトピックから構成される場合のトピック間の比率のことである．著者や発言者によって任意のトピックが選択され，トピック分布として配分されると考えると離散連続モデルを用いたモデリングが着想可能である．

離散連続モデルとは，離散的な選択行動と連続量に関する選択行動が共通の因子で関連付けられている状態を記述するモデルである．離散連続モデルについては，福田・力石 [6]によって詳細にレビューされている．先述のとおり，トピック分布は 1 つの文書が複数のトピックから構成される状態を示している．この場合，複数個の離

散選択肢を同時に選択し，かつ，選択した選択肢の量を配分問題として扱うモデルが必要である．このようなモデルに Multiple Discrete Continuous Extreme Value (MDCEV)モデル [7]がある．MDCEV モデルの適用研究として，北村ら [8]は，活動パターンの選択と活動時間の配分を同時に行うモデルを構築し，多項ロジットモデルとの比較を行っている．また，Jian ら [9]は，カーシェアリング車種を選択し，予算を配分するモデルを構築することで，車両利用パターンに影響を与える因子について分析している．しかし，MDCEV モデルを自由記述の分析に活用した例は，筆者の知る限り存在しない．

以上の背景を踏まえ，本研究では，トピックモデルと離散連続モデルを組み合わせた自由記述の新たな分析フレームを提案し，その有用性を実証的に検証することを目的とする．提案する分析フレームにより，文書中のトピックの抽出，ならびに，抽出した話題に言及する個人属性を客観的に精査することが可能となる．

2. 既往研究のレビュー

本章では自由記述の量的分析に焦点を当て，既往研究を整理する．土木計画・都市計画分野における自由記述の分析では，ワークショップや委員会の討議録やアンケート調査中の自由回答項目，SNS などが対象となる．これらの中でも，特に研究事例が多いのは，討議録に関する分析である．

藤澤ら [10]は，仮想的な討議実験を行い，会話分析のコーディングによって討議中の意見を分類している．また，分類された意見を時間と発言者を軸にとった平面上に布置し，討議の展開などの討議プロセスを可視化を行っている．安藤ら [11]は，道路課金施策のグループインタビューを対象に，質的なコーディングを利用し，グループインタビュー参加者の賛否態度特性を分析している．また，自己組織化マップを討議段階ごとに適用し，討議経緯を把握している．難波ら [12]は，公共交通事業に関する委員会の発言録から MI 値による共起関係により話題と主要な発言，意見を定義し，マルコフ推移確率を用いて意見推移，発言者推移の把握を行っている．岩見ら [13]は，河川整備計画に関する委員会の討議録を対象に，ネットワーク分析の手法を適用し，委員間の協調・対立関係を把握している．また，クラスター分析により，委員の発言テーマの傾向を分析している．福井ら [14]は，地域住民と行政職員との議論から，GTA に基づくコーディングによりトピックとファセット情報を抽出している．また，抽出されたトピックとファセット情報の推移をテキストマイニングの手法により分析することで，コミュニティの成長プロセスを明らかにしている．討議録に関する分析では，サンプル(発言者)が概ね 10 人程度と少ないが，討議プロセスが長期に渡ることや対話という複雑なコミュニケーションを扱うことから，トピックの抽出や変遷の把握などの討議の構造化・討議プロセスの可視化，または発言者の賛否態度などのファセット分析が主な関心となっている．しかし，抽出したトピックと発言者の属性を精査した研究事例は見当たらない．

討議録とは異なるが，この抽出したトピックと発言者の個人属性に着目した研究に，岩見ら [15]の研究がある．岩見らは，パブリックコメントを対象に，クラスター分析によって論点を特定している．また，論点と年齢層のクロス集計と残差分析によって，論点と各論点への言及傾向を把握している．

以上のように，土木計画・都市計画分野の自由記述の研究は積み重ねられてはいるが，トピックとそれに言及した発言者の属性の関係を分析したものは数少ない．本研究で提案する分析フレームは，非集計分析により精緻な分析が可能な点，同時に複数の個人属性を用いた分析ができる点で有用であると考える．

3. 分析フレーム

本研究で提案する分析フレームは，トピックモデルと離散連続モデルにより構成される．まず，事前処理を行ったデータにトピックモデルとしてLatent Dirichlet Allocation[16](以下，LDA) を適用し，本調査の自由記述中に出現した話題を抽出する．また，発言者ごとのトピック分布を算出し，発言者ごとの話題の確率分布を求める．次に，この発言者ごとのトピック分布を，発言者の話題の選択・配分行動の結果とみなし，離散連続モデルの一種であるMDCEVモデルを用いたモデル化を行う．

(1) 分析の事前処理

語数を算出するために形態素解析を行う．形態素解析器として「MeCab」[17]を用いる．形態素解析とは，文章を意味のある単語に区切り，辞書を利用して品詞や内容を判別することである．例えば，「解体を早くしたい」であれば，「解体」「を」「早い」「する」「たい」のように5つの語に区切ることができる． この中から，本分析では出現頻度3回以上の名詞，形容詞，動詞(非自立語，接尾語，数を除く)のみを利用する．

文書内の語の特徴量の尺度として tf - idf 値を用いる．tf 値(ターム頻度)は一文書内での語の出現しやすさを示し，idf 値(文書頻度の対数)は語が特定の文章中で集中して用いられることを表す．よって，これらの積で表される tf - idf 値は，その値が大きい語ほど文書内での重要度が高いことを示す．以下に ti - idf 値の算出法を示す．

$$\text{tf-idf}_{d,w} = \frac{n_{w,d}}{N_d} \times log\frac{|D|}{df_w} \quad (1)$$

$n_{w,d}$: 単語 w の文書 d における出現回数
N_d : 文書 d における総語数
D: 総文書数
df_w : 単語 w が出現する文書数

後述の Bag of Words(BoW)では，この値を語の重みとし，出現回数との積をとることとする．

(2) トピックモデルによるトピックの抽出

トピックの抽出には，トピックモデルの一種である LDAを用いる．ここで，LDA の概要について説明する．なお，LDA についての解説は参考文献[16]に基づいて示す．

LDA はBag of Words(BoW) 表現された文書集合を生成するための確率モデルである．BoW表現とは，文章中に現れる単語のベクトル表現である．また，BoW表現は文章の構造は無視しており，単語の出現回数と共起性を表している．LDAは，BoWから得られる単語の共起性を用いて単語と文書をクラスタリングする手法として用いられる．

LDA では，文書中の各単語に，BoWからは直接得ることのできない潜在変数（トピック）を仮定する．また，LDAの特徴として，文書は複数のトピックから構成され，トピックの構成比としての確率分布をもつ．具体的には，文書 d の i 番目の単語を w_{di} として，対応する潜在変数を z_{di} と定義する．ここで，トピック数を K とし，$\theta_{d,k}$ (k = 1, 2, …,K)を文章 d でトピック k が出現する確率とする．トピック分布は $\theta_d = (\theta_{d,1}\ldots, \theta_{d,K})$ となる．また，各トピックはそれぞれに対応した単語の出現分布 ϕ_k (k = 1, 2, …,K) を有している．文書数を D，文書 d の文章長（総単語数）を N_d とする． ϕ_{dv} をトピック k における単語 v の出現確率とし，単語の出現分布を $\phi_k = (\phi_{k1}\ldots, \phi_{kV})$ とする．

θ_d や ϕ_k はDirichlet分布による生成を仮定をするので，以下のように整理できる．

$$\theta_d \sim \text{Dir}(\alpha), d = 1,\ldots, M \quad (2)$$
$$\phi_k \sim \text{Dir}(\beta), k = 1,\ldots, K \quad (3)$$

ここで，ハイパーパラメータ α, β はそれぞれトピック数 K，単語数 V の次元をもつ．潜在トピックと各単語は以下のように生成される．

$$z_{d,i} \sim \text{Multi}(\theta_d), i = 1,\ldots, n_d \quad (4)$$
$$w_{d,j} \sim \text{Multi}(\phi_{z_{d,i}}), i = 1,\ldots, n_d \quad (5)$$

(3) MDCEVによるトピックの選択・配分のモデル化

総トピック数 K 個のうち，文書 d において j 個のトピックが選択されるとする．この選択行動によって得られる総効用は U_{dk} となり，j 個のトピックは制約条件内で効用 U_{dk} が最大化するように配分されると仮定する．ここで，制約条件は，文書 d におけるトピック分布 $\theta_{d,k}$ の総和であるため，以下のようになる．

$$\text{subject to} \quad \sum_{k=1}^{j} \theta_{d,k} = 1 \quad (6)$$

効用最大化問題を以下の通り設定する．

$$\text{maximize} \quad U_d = \sum_{k=1}^{j} u_{d,k} \quad (7)$$

ここで，$u_{d,k}$ は文書 d においてトピック k を配分することによって得られる効用であり，以下の式により定義する．

$$u_{d,k} = \sum_{k=1}^{j} \frac{\gamma_k}{\alpha_k} \Psi_k \left\{ \left(\frac{\theta_{d,k}}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right\} \quad (8)$$

$$\Psi_k = \exp(\beta x_d + \varepsilon_k) \quad (9)$$

ここで，ψ_k はトピック k に対する基準選好関数であり，配分確率が 0%のときの限界効用に一致する．この基準選好関数の大小により，各トピックへ配分される確率の大きさが決定される．β はパラメータ，x_d は文書 d の著者・発言者に関する個人属性などの説明変数である．また，ε_k はガンベル分布に従う誤差項である

ここで，式(7)の最大化問題に Kuhn-Tucker 条件を 適用することにより，K 個のトピックから j 個のトピックを選択し，それぞれに $\theta_{d,2}$ から $\theta_{d,j}$ のトピック分布確率を配分する場合の選択確率は以下の式が得られる．

$$\begin{aligned} &P(\theta_{d,2}, \theta_{d,3},\ldots,\theta_{d,j}, 0, 0, \ldots, 0) \\ &= \left[\prod_{k=1}^{j} \frac{1-\alpha_k}{\theta_{d,j}+\gamma_k}\right] \left[\sum_{k=1}^{j} \frac{\theta_{d,j}+\gamma_k}{1-\alpha_k}\right] \left[\frac{\prod_{k=1}^{j} e^{V_d}}{(\sum_{k=1}^{K} e^{V_d})^j}\right] (j-1)! \end{aligned} \quad (10)$$

V は Kuhn-Tucker 条件における効用の確定項である．

ここで，α と γ は同時に推定できないため，$\alpha = 0$ と固定して推定を行う．よって式(8)と式(10)はそれぞれ以下のように書き換えられる．

$$u_{d,k} = \sum_{k=1}^{j} \gamma_k \Psi_k \left(\frac{\theta_{d,k}}{\gamma_k} + 1 \right) \quad (11)$$

$$\begin{aligned} &P(\theta_{d,2}, \theta_{d,3},\ldots,\theta_{d,j}, 0, 0, \ldots, 0) \\ &= \left[\prod_{k=1}^{j} \frac{1}{\theta_{d,j}+\gamma_k}\right] \left[\sum_{k=1}^{j} \theta_{d,j} + \gamma_k\right] \left[\frac{\prod_{k=1}^{j} e^{V_d}}{(\sum_{k=1}^{K} e^{V_d})^j}\right] (j-1)! \end{aligned} \quad (12)$$

以上，モデルの定式化を行った．

4. 分析の対象

分析対象とするのは，仮設住宅以降のお住まいについての意識調査(以下，益城町仮設住宅聞き取り調査)である．本調査は，益城町の仮設住宅居住者を対象に，訪問面接調査として行われた．調査目的は以下の 2 点である．

- 町民が必要とする災害公営住宅の戸数，希望する場所などの把握

・　現時点での不自由な点，不安などの幅広い把握

本調査における自由回答形式の設問は，「益城町の復興計画を作るにあたっての意見や要望」と「行政，大学などへの意見・要望も含めて，現在の気持ち・心境」の 2 点である．表-1 に調査票中の自由回答項目を示す．本研究ではこの両方を分析対象とする．また，ここには設問の答えに該当しない回答者の発言も極力全て自由回答として記入されており，これも分析の対象となっている．

本調査は，2016 年 6 月 30 日から同年 11 月 20 日までに行われた．この期間内に，1196 戸の調査を完了した．これは分析対象の 17 団地中，未入居世帯を除いて 81.4%の実施率となる．なお，調査票や基礎集計などは益城町復興計画の資料等[18][19]を参考されたい．

なお，自由回答の有無に関する定義などは，筆者ら[20]の先行研究に従っている．なお，この先行研究は，対話時間と対話中の単語数を推定するモデルを構築し，両モデルの比較から，間接的にトピック量と個人属性の関係を考察したことに相当する．本研究は，トピックモデルにより得られたトピックを被説明変数としたモデルを構築する．本研究のモデルでは，トピックと個人属性の関係を直接記述し，また，トピックの内容別の考察が可能である点が優れている．

5.　分析結果と考察

(1)　トピックの抽出結果

サンプリング方法には，崩壊型ギブスサンプリングを用いた．また，ハイパーパラメータをα=0.5，β=0.5，サンプリング数は5000回とし，トピック数は10とした．表-2にトピック毎に分類された上位10語とトピックタイトルを示す．このトピックタイトルは，分類された単語をもとに，これらの文章中での用いられ方も考慮してトピックのラベル付けを行ったものである．具体的には，「バリアフリー」「避難期・家族の様子」「情報」「隣人との関係」「インフラ整備」「解体・がれき処理」「交通・アクセス」「住まいの再建」「仮設の用地不足」「行政」となっている．

図-1にはトピック割合の平均を示した．「避難期・家族の様子」が最も高くなっている．次いで割合が高くなっているのは「交通・アクセス」「インフラ整備」「仮設の用地不足」「バリアフリー」となっている．

(2)　MDCEVモデルによる分析

最も平均分布の大きいトピックである「避難期・家族

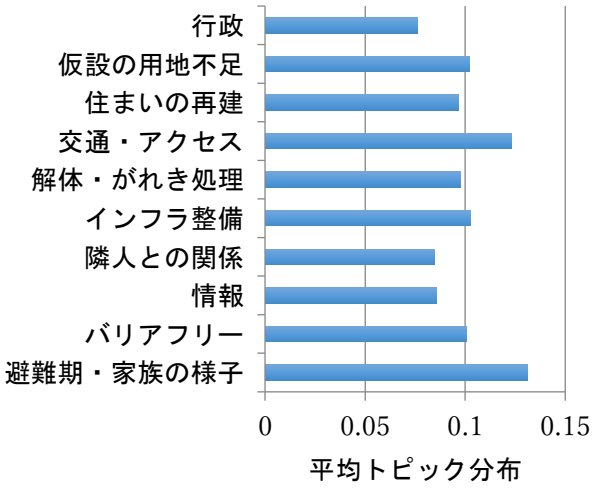

図-1　トピック別の平均トピック分布

表-1　益城町仮設聞き取り調査の自由回答項目

番号	質問	形式
問6	益城町の将来についてお聞かせください	
	(1) 益城町の復興・復旧において，重要と思う点をお聞かせください(複数選択可) 1.生活再建　2.災害に強いまちづくり　3.仕事場の確保　4.コミュニティの維持・確保 5.災害瓦礫の処理　6.子供の教育環境の改善　7.保険・医療・福祉の体制強化 8.情報提供・相談体制の充実　9.その他(次の質問でお答えください)	選択形式
	(2) 益城町の復興計画を作るにあたっての意見や要望等がありましたら教えてください	自由記述
問7	行政，大学などへの意見・要望も含めて，現在のお気持ち・心境をお寄せください	自由記述

表-2　トピックの抽出結果

トピックタイトル	バリアフリー	避難期・家族の様子	情報	隣人との関係	インフラ整備	解体・がれき処理	交通・アクセス	住まいの再建	仮設の用地不足	行政
1	付ける	避難	情報	気	道路	解体	バス	2年	スペース	聞く
2	玄関	行く	掲示板	音	地盤	公費	買い物	土地	置く	災害
3	風呂	息子	知る	知り合い	戻る	何	遠い	災害公営住宅	荷物	まちづくり
4	悪い	自宅	近所	違う	県道	時間	バス停	住宅	子供	必要
5	高い	車中泊	電話	嬉しい	急ぐ	自分	団地	建てる	駐車	支援
6	手すり	全壊	集会	騒音	解体	処理	不便	住める	駐車場	要望
7	スロープ	暮らす	良い	見る	整備	遅い	交通	再建	場所	思う
8	トイレ	住む	入る	カビ	早い	手続き	欲しい	確保	雨	希望
9	段差	いる	回覧板	避難所	進める	いつ	移動	先	洗濯	強い
10	目	実家	回る	いる	自宅	がれき	便	金	子供達	意見

の様子」を選好基準として，MDCEVモデルによるパラメータ推定を行った．サンプル数は，説明変数の欠損データを除いたため，N=921である．推定結果を表-3に示し，トピック毎の結果と考察を以下に述べる．

「バリアフリー」では，仮設住宅入居日数と自分専用・家族共用車有ダミーがパラメータが負で有意となった．ここから，仮設に入って間もないか，自分専用もしくは家族共用の車がないとこのトピックに触れる傾向にあることがわかる．仮設のバリアフリー化は早期に解決されるべき課題であることなどが考えられる．また，トピック「バリアフリー」は高齢者の言及が多いトピックであると予想されたが，年齢に関する変数は本モデルでは有意傾向になかったことから，説明変数から除いている．

「情報」では，2016 年 6 月 30 日の聞き取り調査開始日からの経過日数と問 6(1)で”8.情報提供・相談体制の充実”を選択した場合のダミー変数を説明変数として投入している．両変数共に有意となっており，聞き取り調査実施日が調査開始直後であるほど，また，問 6(1)で”8.情報提供・相談体制の充実”を選ぶとこのトピックに触れる傾向にあることがわかる．

ここで，問 6(1)と問 6(2)・問 7 の関係を述べる．問 6(1)とは，表-1 に示す通り，自由回答項目(問 6(2)・問 7)直前の選択形式の設問である．例えば，表-1 より，問 6(1)の選択肢に，“9.その他(次の質問でお答えください)”とあるように，聞き取り調査では，問 6(1)で選んだ選択肢に関して問 6(2)・問 7 の自由回答で深く掘り下げて聞き取っている．したがって，問 6(1)の回答と問 6(2)・問 7 の自由回答中のトピックに関連があると考えられる．トピック「情報」における推定結果は，以上の問 6(1)の回答と問 6(2)・問 7 の関連を示しており，本分析手法の有用性を示す一つの要素であると考える．

「隣人との関係」では，仮設住宅の入居日数と問 6(1)で”4.コミュニティの維持・強化”を選択した場合のダミー変数を説明変数として用いた．仮設住宅への入居日数はパラメータが負で有意となっており，仮設入居期間が短いとこのトピックに触れる傾向にあることがわかる．一方，問 6(1)”4.コミュニティの維持・強化”ダミーは有意とはなっていない．表-2 中のトピック「隣人との関係」には，「知り合い」や「いる」という語が出現している．このことから，トピック「隣人との関係」は仮設団地のの知人の有無に関するものと考えたため，問 6(1)”4.コミュニティの維持・強化”ダミーを利用している．このダミー変数が有意とならなかったことから，トピック「隣人との関係」が，表-2 中の「気」「音」などの語から示されるように「音が気になる」のように周囲との不愉快に関するトピックであることが考えられる．

「インフラ整備」では，説明変数として，男性ダミー，持ち家ダミー，自分専用・家族共用車有ダミー，問 6(1)”4.災害に強いまちづくり”ダミーを設定した．すべての変数が有意となり，男性や持ち家世帯，問 6(1)で”2.災害に強いまちづくり”を選ぶとこのトピックに触れる傾向にあることがわかった．表-2 中より，「道路」「県道」

表-3　MDCEVモデルの推定結果

トピック	説明変数	パラメータ	t値
バリアフリー	定数項	0.25	1.88 *
	男性ダミー	-0.23	-1.31
	仮設住宅居住日数	-0.01	-2.12 **
	自分専用・家族共用車有ダミー	-0.38	-2.58 ***
情報	定数項	-0.21	-1.18
	調査開始日(6/30)からの経過日数	-4.9×10^{-3}	-2.25 **
	問6(1)8.情報提供・相談体制の充実ダミー	0.82	5.66 ***
隣人との関係	定数項	-0.2	-1.58
	仮設住宅居住日数	-0.01	-2.09 **
	問6(1)4.コミュニティの維持．強化ダミー	0.17	1.03
インフラ整備	定数項	-1.14	-4.75 ***
	男性ダミー	0.67	4.39 ***
	持ち家ダミー	0.55	2.77 ***
	自分専用・家族共用車有ダミー	0.29	1.82 *
	問6(1)2.災害に強いまちづくり	0.33	2.28 **
解体・がれき処理	定数項	-1.06	-3.2 ***
	年齢	0.01	1.85 *
	男性ダミー	0.39	2.56 **
	問6(1)5.災害がれきの処理ダミー	0.42	2.94 ***
交通・アクセス	定数項	0.67	1.53
	年齢	-0.01	-1.83 *
	男性ダミー	0.18	1.11
	世帯人数	0.12	1.67 *
	自分専用・家族共用車有ダミー	-0.58	-3.42 ***
	世帯保有自動車数	-0.16	-1.64
	災害公営住宅希望ダミー	0.57	3.72 ***
住まいの再建	定数項	-0.17	-0.99
	男性ダミー	0.3	2.04 **
	調査開始日(6/30)からの経過日数	-2.4×10^{-3}	-1.18
	農家ダミー	0.44	2.14 **
仮設の用地不足	定数項	1.06	3.69 ***
	年齢	-0.02	-3.44 ***
	仮設住宅居住日数	-3.8×10^{-3}	-1.3
	津森・平田団地ダミー	-0.53	-1.86 *
	非就業者ダミー	-0.19	-1.32
行政	定数項	-0.19	-0.63
	年齢	-0.01	-1.24
	男性ダミー	0.52	3.25 ***
	対話時間(分)	-1.7×10^{-4}	1.15
サンプルサイズ	921		
最終尤度	-4595		

注)***1%有意，**5%有意，*10%有意

が抽出されていることから，このトピックは，主に道路整備に関するトピックであると考えれれる．ここから，自動車を保有する回答者がこのトピックに触れやすいのではないかと考え，自分専用・家族共用車有ダミーを投入した．この変数が統計的に有意になったことから，ドライバー視点から道路整備の声が多かったことが予想される．その他，トピック「情報」と同様に選択形式の回答と自由回答の関連も示されている．

「解体・がれき処理」では，年齢，男性ダミー，問 6(1)”5.災害がれきの処理”ダミーを説明変数として利用した．推定結果から，高齢，男性，問 6(1)で”5.災害がれきの処理”を選ぶとこのトピックに触れる傾向にあることが明らかとなった．このトピックでもトピック「情報」や「インフラ整備」と同じく選択形式の回答と自由回答の関連があることがわかった．

「交通・アクセス」では，年齢，世帯人数，自分専用・家族共用車有ダミー，災害公営住宅希望ダミーを説明変数として利用している．推定結果より，若年層，世帯人数が多い，自分専用または家族共用の車がない，仮設後の住まいに災害公営住宅を希望しているとこのトピックに触れる傾向にある． 高齢者ほど交通・アクセスに関する言及が多いと予想されたが，予想に反して年齢が若いほうが言及しやすいという結果になっている．一方で，自家用車がないとこのトピックに触れる傾向にあるという仮説を立てていたが，このモデルよって統計的に仮説の正しさが検証された．また，災害公営住宅への入居希望者がこのトピックに触れやすいという結果が得られたため，災害公営住宅の建設に際しては，特に交通・アクセスに配慮する必要があることが推察される．

「住まいの再建」では，男性ダミー，2016 年 6 月 30 日の聞き取り調査開始日からの経過日数，農家ダミーを説明変数とした．男性ダミーと農家ダミーが有意となり，男性や農家がこのトピックに触れる傾向にあることがわかった．

「仮設の用地不足」では，年齢，仮設住宅居住日数う，津森・平田団地ダミー，非就業者ダミーを説明変数として設定した．若年層や津森・平田団地居住者でこのトピックに触れる傾向にある．津森・平田団地とは，益城町内でも農村部に位置する仮設団地である．

「行政」では，男性ダミー，年齢，調査中の対話時間を説明変数とし，男性の方がこのトピックに触れる傾向にあることを明らかにした．

以上の結果をまとめると，「インフラ整備」と「行政」のような行政に関する話題や，「解体・がれき処理」と「住まいの再建」などの住宅の再建に関する話題は，男性の方が関心を持ちやすい話題であると言える．また，問 6(1)の選択形式設問の回答と問 6(2)・問 7 の自由回答中の関係が有意となったのは，「情報」「インフラ整備」「がれき処理」の 3 つであった．問 6(1)の選択形式の設問のうち，「情報提供・相談体制の充実」「災害に強いまちづくり」」「災害がれきの処理」は，特に関心のある項目であったとも考えられる．また，先行研究[20]では対応分析を用いて，対話中の単語と個人属性の関係を分析したが，本分析手法では，トピックと個人属性の関係を直接的に，かつ統計的に示すことができた．

6. おわりに

本研究では，トピックモデルと離散連続モデルを統合した新たな自由記述データの分析法を提案した．2016年熊本地震による益城町仮設住宅居住者を対象とした聞き取り調査中の自由回答を事例とした分析の結果から，属性別の回答傾向の違いや，選択式設問の回答と自由回答中のトピック選択の対応が統計的に示され，手法の有用性が確認できた

本分析法は，聞き取り調査の実施が出来なかった回答者の回答トピックを予測することも可能である．聞き取りを実施した回答者と聞き取りが未実施の回答者の回答トピックを比較分析することが今後の展開として考えられる．また，本分析法のみでは，例えばトピック「交通・アクセス」が仮設住宅の交通環境に対してポジティブまたはネガティブな文脈で語られていたかは判別できない．そこで，トピックに対する賛否などの態度を分析することにも価値があると考えられる．その他にも既往研究[4]におけるトピック数は本研究で抽出したトピック数より大きくなっていることから，よりトピック数が大きくなるデータに本手法を適用し，有用性を検証することが望まれる．

謝辞：仮設住宅聞き取り調査は，益城町復興課との共同実施によるものです．また青山学院大，慶応義塾大，東京大，自治医科大，関西学院大，京都大，九州大，九州工業大，佐賀大，大分大，鹿児島大，熊本学園大，熊本県立大及び熊本大の学内の多くの皆様にボランティアでご協力をいただきました．深く感謝申し上げます．

参考文献

1) Blei, D. M., Andrew, Y. N. and Michael I. J.: Latent dirichlet allocation, Journal of Machine Learning Research 3, 2003.
2) Lijun Sun, Yafeng Yin：Discovering themes and trends in transportation research using topic modeling, Toransportation Resaerch Part C: Emerging Techinologies, Vol.77, pp.49-66, 2017.
3) 塚井正人，原祐輔，山口敬太，大西正光：土木計画学の研究トピックスの変遷，第 56 回土木計画学研究

発表会・講演集，Vol.56，2017.
4) 塚井誠人，椎野創介：討議録に対するトピックモデルの適用，土木学会論文集 D3,Vol.72,No.5,2016.
5) 神谷啓太，布施孝志：トピックモデルを利用した地域別人口特性の把握手法の提案，第 55 回土木計画学研究発表会・講演集，Vol.55，2017.
6) 福田大輔，力石真：離散-連続モデルの研究動向に関するレビュー，土木学会論文集 D3(土木計画学)，Vol. 69，No. 5，pp. I_497-I_510，2013.
7) Bhat, C.: The multiple discrete-continuous extreme value(MDCEV) model: role of utility function parameters,identification considerations, and model extensions, Transportation Research Part B: Methodological, Vol. 42, No. 3,pp. 274–303, 2008
8) 北村拓也，栁沼秀樹，寺部慎太郎，康楠：活動パターンと時間配分の同時選択を考慮したアクティビティモデルの構築，第 55 回土木計画学研究発表会・講演集，Vol.55，2017.
9) Sisi Jian , Taha Hossein Rashidi, Vinayak Dixit: An analysis of carsharing vehicle choice and utilization patterns using multiple discrete-continuous extreme value (MDCEV) models, Transportation Research Part A: Policy and Practice, Vol.103, pp.362-376, 2017.
10) 藤澤徹，秀島栄三，北村直之：地域社会の課題解決に向けた住民討議プロセスに関する実験的分析，社会技術研究論文集，Vol.5, 88-95，2008
11) 安藤章，森川高行，三輪富生，山本俊行：討議過程の可視化手法を用いた道路課金政策に対する市民の賛否態度特性の分析，都市計画論文集，No.45-3，pp.481-486，2010.
12) 難波雄二，塚井誠人，桑野将司：文脈マイニングモデルを用いた討議過程の可視化手法に関する研究，土木学会論文集 D3(土木計画学)，Vol.67, No.5, pp. I_209-I_219，2011.
13) 岩見麻子，大野智彦，木村道徳，井手慎司：公共事業計画策定過程の議事録分析による意見の協調・対立関係把握のための分析手法の開発，土木学会論文集 G（環境），Vol.70，No.6，pp.II_249-II_256，2014.
14) 福井のり子，力石真， 藤原章正：農村地域の活性化にむけた初動期における個人とコミュニティの成長プロセスグラウンデッド・セオリー・アプローチ（GTA）と複線経路・等至性アプローチ（TEA）による検証，都市計画論文集，Vol.52，No.2，pp.209-219，2017.
15) 岩見麻子，宮下知己，井手慎司：大規模パブリックコメントの論点把握に対するテキストマイニングの有用性の検討，土木学会論文集 G（環境），Vol.71，No.6，pp.II_13-II_21，2015.
16) 岩田具治：トピックモデル，講談社，2015.
17) MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/ (2018 年 2 月閲覧)
18) 熊本県益城町: 益城町復興計画, pp. 99-111, 2016.12
19) 渡邉萌，佐藤嘉洋，円山琢也：熊本地震の復興初期における益城町仮設住宅入居者の居住地選択意向,都市計画論文集,Vol.52,No.3,pp.1094-1100,2017
20) 川野倫輝，佐藤嘉洋，円山琢也：対話時間と単語数を考慮した聞き取り調査の自由回答分析方法の開発提案- 熊本地震における益城町仮設住宅聞き取り調査への適用 -，都市計画論文集，Vol.53，No.1，2018

(2018. . 受付)

INTEGRATING TOPIC MODEL AND DISCRETE-CONTINUOUS MODEL TO ANALYZE FREE ANSWER DATA

Tomoki KAWANO, Yoshihiro SATO and Takuya MARUYAMA

Text mining approach for analyzing free answer data attracts attention in infrastructure and city planning studies. This paper proposed a novel framework integrating topic model and discrete-continuous model to analyze free answer data. We use the interview survey data for households in Mashiki temporary housings following the 2016 Kumamoto earthquake. Topic model describes the ratio of topic in interview for each respondent and discrete-continuous model describes the topic choice and the ratio of topic. Our new framework statistically demonstrates the relationship between respondent attributes and topic in interview.