2　Twitter 情報分析による土砂災害の前兆現象等情報の収集・把握手法の検証

2.1　Twitter 情報の収集・把握手法

2.1.1　Twitter データ

(1) 基本事項

Twitter は，米国の Twitter 社が提供する，Web 上のマイクロブログサービスのひとつである。Twitter の利用者は，ツイートと称する 140 文字の短文を Web 上に投稿することができ，投稿されたツイートは，投稿者が制限を掛けない限りにおいては，一般に公開される。全世界で，3 億 2000 万人の月間アクティブユーザー[1]を抱える巨大な Web サービスであり，国内でも広く利用されている。また，Twitter 上の情報は，Web ブラウザやスマートフォンアプリを利用して閲覧することができる他，Twitter 社が提供する Web 上の Application Programming Interface（以下，「API」という）を用いて取得することも可能である。したがって，ソーシャルメディア情報として，2 次的な活用を考えることができるものである。一方，匿名で利用できるサービスであることから，必ずしも正しい情報が流通しているとは限らないことに注意する必要がある。

Twitter の情報としての特徴は，以下の 2 点である。

- **投稿される情報にリアルタイム性がある**

  ツイートは 140 文字以内に制限されており，利用者は，長文を投稿することができない。この制約から，利用者は通常のブログなどと比較して，自分自身の意見や観察した情報を即座に投稿する傾向がある。

- **情報の拡散が早い**

  リツイートと呼ばれる，他人の発言を複写して簡単に再投稿する機能を有している。利用者は，リツイート機能を用いて，他の利用者が投稿したツイートに対して，関心を持ったり，何かしらの意見を持ったりした場合に，当該ツイートを簡単に投稿することができる。このため，Twitter は，他のソーシャルメディアと比較して，インパクトのある情報が拡散しやすい傾向にある。一方，ツイートデータに対して統計的な処理を行う場合には，リツイートによって同じ情報が複製されて水増しされることを考慮する必要がある。

(2) データの収集方法

Twitter 上に投稿されたツイートを，データとして収集するための手段としては，大きく分けて，(a)無償公開されている Web 上の Twitter API を利用する方法と，(b)Twitter 社のデータを販売する業者を通じて有償で購入する方法がある。

(a)のメリットは，無償でデータを入手できることと，Twitter API へのアクセス処理を工夫することで，リアルタイムに近い形でデータを収集できることである。一方，デメリットとしては，Twitter 上に存在するツイートデータを，もれなく入手できるわけではないこと

[1] Twitter 社の日本国内 Web ページより引用。https://about.twitter.com/ja/company

である。Twitter API は，用途に応じて複数用意されており[2]，代表的なものとして，検索キーワードを指定して当該キーワードを含むツイートを収集できる「Search API」がある。この API は，指定したキーワードを含むツイートを，できる限り提供するものである。無償であることから，本 API を活用したスマートフォンアプリや，Web サービスが数多く存在している。しかしながら，Twitter 社が定める取得上の制限[3]があり，必ずしもすべてのデータを取得できるわけではない。また，キーワードを指定する方式であるため，取得したいツイートに含まれると想定されるキーワードを，あらかじめ調査して確定させておく必要がある[4]。

一方，(b)による方法としては，大きく分けて，i)日付やキーワードを指定して条件に該当するデータを一括で購入する方法と，ii)随時リアルタイムに全量データを入手できる契約を結ぶ方法がある。i)は，日付やキーワードを指定することで，当該条件を満足する全ツイートを入手できることが特徴だが，リアルタイム性は期待できない。このため，ケーススタディのように，過去の何らかの事例に対してデータを俯瞰的に調査する用途に適している。本研究においても，一部の調査を行うために，過去データの購入を実施した。一方，ii)の方法は，非常に高コストであり，主に大規模な Web サービスを展開する企業が話題分析やマーケティングに利用されていることが考えられる。このため，研究用途としては利用されていないものと想定される。

2.1.2 関連研究

Twitter に代表されるマイクロブログに対するデータマイニング技術に関しては，奥村（2012）[13]にて概説されている。この中で，ソーシャルメディアの情報から特定の話題を検出する技術として，Kleinberg（2002）によるバースト検知[14]が紹介されている。Kleinberg は，Web 上で通常ランダムに出現する特定のキーワードが，ある時に急増する現象を「バースト」と呼んだ。このバーストを捉えることにより，特定のキーワードに関連する話題やイベントの発生を捉えることができるという仮説がある。

高橋（2011）は，Twitter 上に投稿されるツイートは，実世界で起きている事象を捉えるためのセンサ（ソーシャルセンサ）であると仮定し，Twitter データを用いて花粉症の広がりを捉える手法を提案している[15]。

これらの研究成果を踏まえ，著者らは，防災分野への適用を検討するため，Twitter を用いて災害の発生をいち早く捉えるための技術を開発した[16][17]。この研究により，浸水災害や土砂災害といった事象においては，Twitter からリアルタイムに発災情報を捉えることができる可能性があることがわかった。

---

[2] Twitter 社の開発者向け Web ページを参照。https://dev.twitter.com/rest/public

[3] Twitter の Search API に対して通常のユーザー認証方式にてアクセスした場合,15 分に 180 回までアクセス可能であり，1 回のアクセスで標準 15 件，最大 100 件まで取得可能となっている。これらの仕様については，今後予告なく変更される可能性がある。

[4] 富士通研究所においては，研究用途での利用を前提として，2012 年 7 月から災害や防災に関連するキーワードを含むツイートを収集し，データ調査や研究開発に活用してきた。

2.1.3 本研究のアプローチ

本研究では，土砂災害が発生する前の警戒に資する情報を Twitter データから抽出することを目的として，データの調査と手法の検討を行った。本研究のアプローチを以下に示す。

1. **特定の災害事例に対する詳細なデータ調査（初期調査）による災害発生状況把握の可能性検討**
   特定の災害事例を対象として，気象条件や災害の発生事実と Twitter データを突き合せることにより，災害の発生前後でどのようなツイートが投稿されているか調査し，情報の特性と利活用の可能性を考察した。（2.2，2.3 節）
2. **俯瞰的なデータ調査による警戒期から発災前における状況把握の可能性検討**
   土砂災害の発災前後の Twitter データの全体像を掴み，Twitter データから抽出できる可能性がある情報を整理するため，統計解析手法を用いた俯瞰的なデータ調査を実施した。（2.4，2.5 節）
3. **前兆現象等情報の収集・把握のための統計手法の検討**
   以上の調査結果を踏まえ，土砂災害が発生する前の警戒的な情報を Twitter データから自動的に取得するための手法と課題を考察した。（2.6 節）

2.2　災害事例におけるつぶやきの内容と豪雨・災害事象との関係

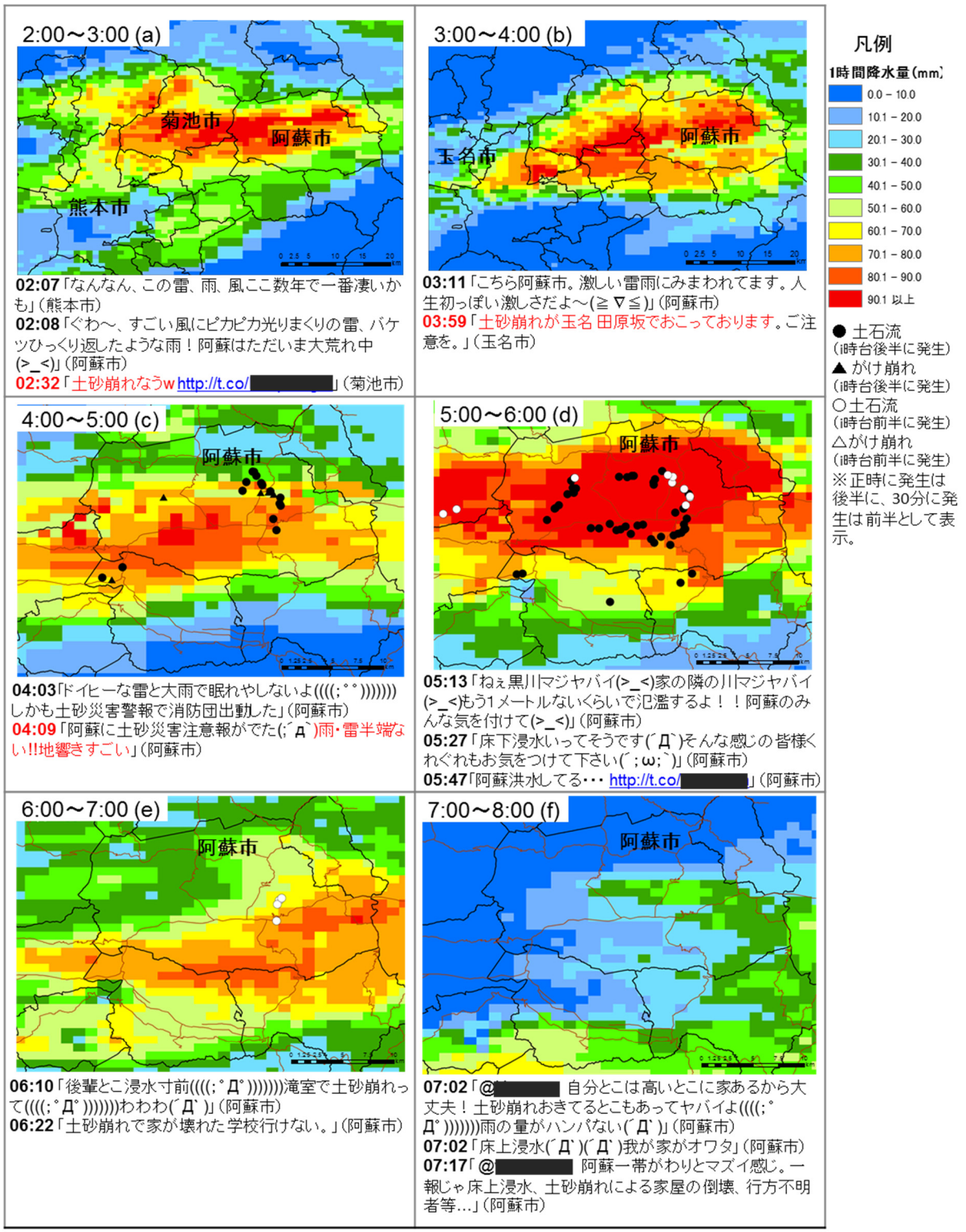
2.2.1　平成 24 年 7 月九州北部豪雨

平成 24 年 7 月九州北部豪雨は，阿蘇地域（阿蘇市・南阿蘇村・高森町）において，土砂災害が 85 件発生し,死者･行方不明者 25 名をもたらした甚大な被害が発生した災害である[18)]。

本研究では，武田らが構築した分析プロセス[17)]を用い，さらに詳細に災害時の分析を進めるため，まず熊本県内全域に大雨注意報が発令された直前の 2012 年 7 月 11 日 16 時 0 分から翌 12 日 12 時までに日本全国で投稿された Twitter データ約 2,850 万件を取得した。これらを分析した結果，熊本県内で投稿されたと推定されるツイートは約 18 万件あり，その内市町村推定までできたものは約 2.4 万件であった。

県内でツイートされたと推定されたもののうち，「大雨」「豪雨」もしくは「土砂降り」を含むツイート数は約 4,100 件，「土砂崩れ」もしくは「土石流」を含むツイート数は約 600 件あった。図-2.1 は，降雨強度（mm/h）の空間分布の変化に伴う土砂災害発生箇所（熊本県から国土交通省に報告のあったもので，人家等に影響のなかったものは除く）の変化と各時間帯においてつぶやかれた主なツイートを示したものである。

当該災害の際，阿蘇市・南阿蘇村に土砂災害警戒情報が発表されたのは 7 月 12 日 2 時 40 分であった。この時の Twitter データを見てみると，その直前の 2 時 32 分に阿蘇市に隣接する菊池市において既に小規模な土砂崩れ（添付されていた写真により確認）が発生していたことを示すツイートがあることが分かる（図 2-1（a））。また，3 時 59 分には玉名市・熊本市（田原坂）で土砂崩れが発生したことをうかがわせるツイートがある（図 2-1（b））。
阿蘇地域において，土石流等が集中的に発生したのは 4 時台後半から 5 時台にかけてであった（図 2-1 (c)及び(d)）。このとき，仮に Twitter データがリアルタイム分析できていれば，近隣地域での前兆現象を，発災の約 30 分から 2 時間前に把握することができ，より早い時間帯に避難指示等を発令することができた可能性がある。



※降水量分布は解析雨量(気象庁提供)を元に作成
※土砂災害発生分布は国土交通省提供のデータを元に作成(発生時刻6時とされている土砂災害は、5:00～6:00の図中で表記)
※括弧内の市町村名は研究チームが推定した結果であり、100%の確かさを保障するものではない

図 2-1　降雨強度（mm/h）の空間分布の変化に伴う土砂災害発生箇所の変化と各時間帯においてつぶやかれた主なツイート

ただし，酒井ら（2013）は，土砂災害警戒情報が発表された段階で，既に猛烈な豪雨が降っており，仮に早期に避難指示が発令されていたとしても，現実的には避難行動をとること

は困難な状況にあったと考察している[19]。このようなケースの場合，前兆現象に関する情報を得られたとしても，住民に対し実際どのような行動を促すかは課題として残る。

土砂災害が集中的に発生し始める直前の 4 時 9 分には，土砂災害の前兆現象ともとれる「地響き」を含むツイートがある。このツイートは共起語として「雷」を含んでおり，雷による震動を感じ「地響き」として投稿している可能性がある。しかしながら，酒井ら(2013)の調査によると，4 時半頃に「地鳴り」を聞いたとの証言を複数得ており[19]，直ちに雷によるものであると決めつけるのは早計のようにも思われる。他の土砂災害の事例を見ても，「地響き」等は比較的よく投稿されており，「地響き」等の前兆現象に関連するキーワードとしての取り扱いは今後の課題となる。

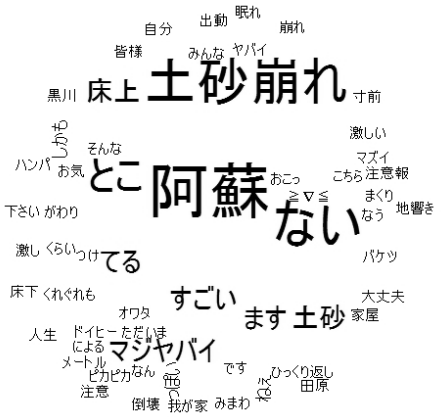


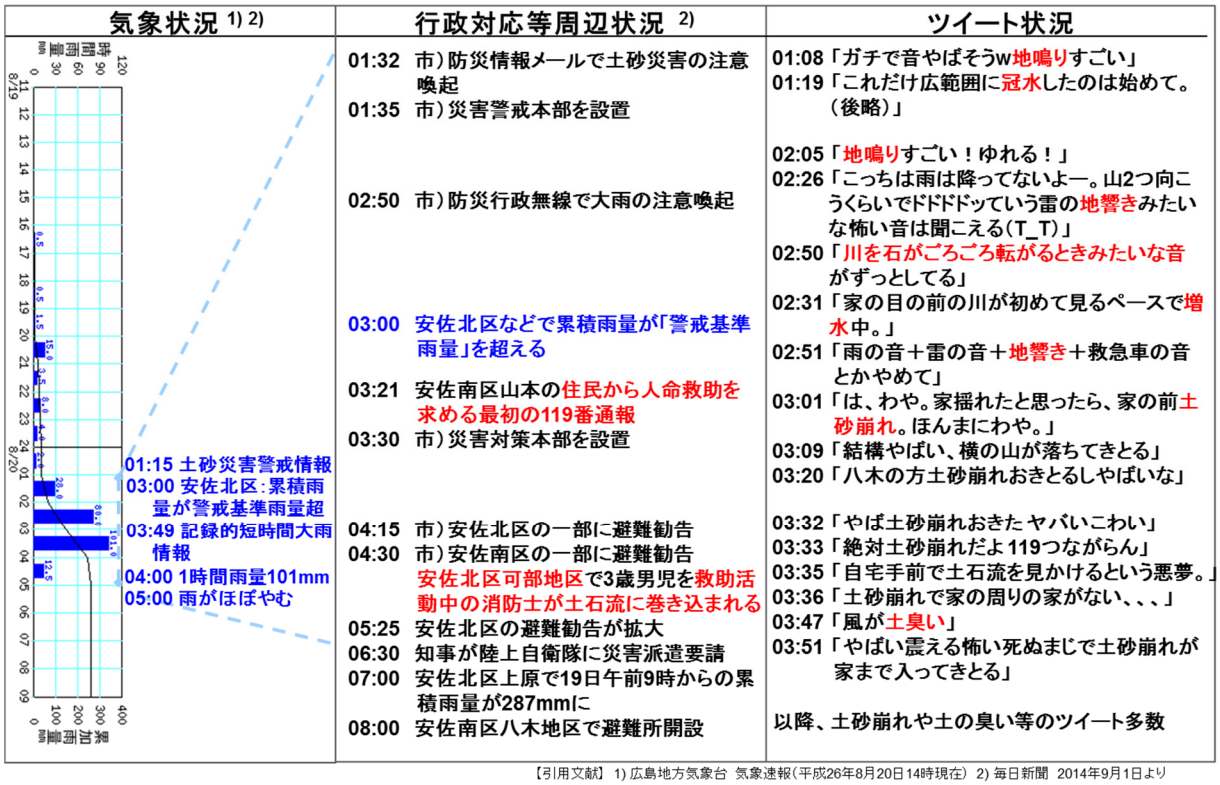
図 2-2　平成 24 年 7 月九州北部豪雨災害（阿蘇地域土砂災害）の際の主なツイートをワードクラウド化したもの
※形態素解析を行い頻度が高い語を大きなフォントで表現

2.2.2　平成 26 年 8 月広島市土砂災害

平成 26 年 8 月 19 日から 20 日未明にかけて広島県広島市で発生した豪雨により，広島市の安佐北区，安佐南区を中心に土砂災害が 167 件発生し[20]，死者 74 名をもたらす甚大な被害が発生した[21]。

8 月 19 日～20 日の気象状況，および 20 日未明の行政対応状況と，筆者らがツイート本文やプロフィール等を見て広島県に関するツイートと判断したものを収集し，時系列に整理した（図 2-3）。災害発生当日，土砂災害警戒情報が発表された 1 時 15 分より以前から，雷を伴う豪雨による「地鳴り」や「冠水」といったツイートが見られている。それ以降，「地響き」や「川を石がごろごろ転がるときみたいな音」などの，土砂災害の前兆現象とも考えられる内容のツイートが投稿されている。そして，災害が集中して発生しはじめたと考えられる 3 時頃を過ぎると，「土砂崩れ」や「土臭い」というツイートが次々に投稿され，未明の外が薄暗い状況の中でも人家周辺に流れ込んだ土砂やその臭いで異常を感じられていたことが推測された。

図 2-3 に記載したツイートのうち，簡易的な方法により投稿場所を推定できるツイートは僅かである。そのため，災害が発生しているおそれがある地域における有用な情報を収集するためには，Twitter ユーザーのプロフィールや過去の発言も利用する方法[22]や，方言など言葉の地域性を考慮した推定を行うことが考えられる。



【引用文献】 1) 広島地方気象台 気象速報（平成26年8月20日14時現在） 2) 毎日新聞 2014年9月1日より

図 2-3　平成 26 年広島災害時の対応等時系列とツイート状況
（筆者らが広島に関するツイートと判断したものを掲載）

2.3　災害発生場の状況把握の可能性検討

前節の事例から，災害時に土砂災害の前兆や発災に関する情報をツイートから得られる可能性があることが分かった。

そこで，九州北部豪雨災害（2012 年 7 月）及び広島豪雨災害（2014 年 8 月）の際のツイートを分析し，Twitter 情報を活用した土砂災害発生場の状況把握の可能性について検討を行った。

2.3.1　分析手法

Web 上に流れる膨大な Twitter 情報からリツイートを除き，ツイートの内容がどこの市町村での出来事を指しているのか絞り込みを行った上で（以下「市町村推定」という。），土砂災害に関連するものを抽出し，その発言内容から状況を把握するため，土砂災害に関連する発言内容をカテゴリー分けし，さらに，各カテゴリーを代表するキーワードの設定を試みた。

設定したキーワードが含まれるツイートを抽出し，各カテゴリーのツイート数の変化に基づく状況の推察及び，状況を直接的に示す単語による，さらに的確な現象の把握の可能性を考察した。なお，市町村推定は武田ら（2014）の手法 [17]を一部簡略化して行った。推定の流れは次の通りである。

① キーワードを設定する。

② 公開されている GPS 情報やユーザープロフィール，ツイート本文中の地名・ランドマー

ク等からユーザーの居住地を推定し，Twitter データを都道府県別に仕分ける。
③ 同種の「つぶやき」の急増をとらえて都道府県単位での異常事態の発生を推定する。
④ 異常事態の発生の推定結果を踏まえ，対象とする都道府県内で投稿されたと推定される Twitter データに絞り込んだ上で②の分析を再度行い，市町村を推定する。

2.3.2 対象とする災害および Twitter 情報の概要

対象災害は，土砂災害発生場の状況を表す諸現象が顕著であり，土砂災害の前兆現象についても住民により把握されていたことが分かっている（例えば，酒井ら，2013）[19]，2012 年 7 月九州北部豪雨（以下「九州北部豪雨」という。）及び 2014 年 8 月豪雨（以下「広島豪雨」という。）によって，熊本県阿蘇市・南阿蘇村，広島県広島市で発生した土砂災害とした。分析対象期間は土砂災害発生前後の時間帯とし，九州北部豪雨では 2012 年 7 月 11 日 16 時～12 日 12 時，広島豪雨では 2014 年 8 月 20 日 1 時～5 時 30 分とした。双方とも，深夜に降雨が強まって明け方にかけて土砂災害が発生し，多数の人的被害等をもたらした。
分析対象ツイートは，分析対象期間に投稿されたツイートのうち，発言者の位置推定ができたもののうち，九州北部豪雨では阿蘇市および南阿蘇村（以下「阿蘇」という。）と推定された 2207 件を，広島豪雨では広島市（以下「広島」という。）と推定された 5813 件とした。

2.3.3 発言内容に基づくカテゴリー分けとキーワード設定

本検討に先立って，「避難勧告等の判断・伝達マニュアル作成ガイドライン」（内閣府，2014）[3]等を参考に土砂災害の現象を示す専門的な用語(表面流発生,小石がぱらぱら落下等)をキーワードとして使用することが可能かどうかを確認したが，ガイドラインに掲載されているような表現のままでツイート中に出現する事例は極めて少なかったことから，実際のツイートの記載内容を読み解いて，改めてキーワードとして設定する方針とした。
設定の流れを以降に示す。
① ツイートの内容に基づき，土砂災害の前兆現象や災害発生に関連するツイートを抽出
② 抽出したツイートを内容からカテゴリーに分類
③ 各カテゴリーを構成する語を抽出
④ 土砂災害に関連する語をキーワードとして設定

ツイートの内容に基づき，広島で 1,617 件，阿蘇で 968 件を土砂災害に関連するツイートとして抽出した。さらに，抽出したツイートを表 2-1 の「状況カテゴリー」に分類した。

表 2-1　状況カテゴリー

項目	内容
気象関係（雨、雷）	雨、雷等天気の状況
警報、避難、消防活動関係	警報発表や避難勧告等の発令、消防活動等の状況
河川関係（増水、流量）	流量の増加等、河川に関する状況
水害関係（氾濫、浸水）	氾濫、浸水等の水害に関する状況
土砂災害関係	土石流、がけ崩れなど土砂災害に関する状況
休校、休業関係	休校、休業などの対応に関する状況
交通関係（通行止め）	通行止め、交通機関の遅延など交通に関する状況
不安感等	災害発生の危険等に対する懸念や感想
その他	上記に分類されないが災害に関する状況

テキスト型データを統計的に分析するソフトウェアである KH Coder を利用し，状況カテゴリーごとに，ツイートに登場する語を品詞分解した上で自動抽出を行った。抽出した語の中から，各状況カテゴリーを表し得る語（例：土砂災害では，「土砂崩れ」「災害」など，不安では，「やばい」「怖い」など）をキーワードとして設定した。土砂災害カテゴリーの例を表 2-2 に示す。ここで，品詞の区分は KH Coder の分類ルールに基づくものであり，「名詞」は漢字を含む 2 文字以上の語を，名詞 C は漢字 1 文字の語を示す。

表 2-2　設定したキーワード例（土砂災害カテゴリー）

九州北部豪雨

抽出語	品詞	出現回数
土砂崩れ	名詞	34
生き埋め	名詞	20
土砂	名詞	8
災害	名詞	5
土石流	名詞	3
崩れ	名詞	3
崩れる	動詞	3
埋まる	動詞	3
がけ崩れ	名詞	1
なぎ倒す	動詞	1
安否	名詞	1
溢れる	動詞	1
壊れる	動詞	1
地響き	名詞	1
倒木	名詞	1
被害	名詞	1

広島豪雨

抽出語	品詞	出現回数
土砂崩れ	名詞	107
土砂	名詞	29
災害	名詞	26
崩れる	動詞	11
裏山	名詞	9
生き埋め	名詞	4
流れ込む	動詞	4
崖	名詞C	3
山	名詞C	3
がけ崩れ	名詞	2
流れる	動詞	2
土砂降り	名詞	1
土石流	名詞	1
被害	名詞	1

表 2-1 の各カテゴリーにツイート内容を読み解いて分類したツイート数（図 2-4「内容分類による抽出」）及び，各状況カテゴリーのキーワードを含むツイート数（図 2-4「キーワードによる抽出」）の時系列変化を比較すると，両者の傾向は概ね一致し，設定したキーワードで各状況カテゴリーのツイート数の変化を捕捉することが確認できた（図 2-4）。

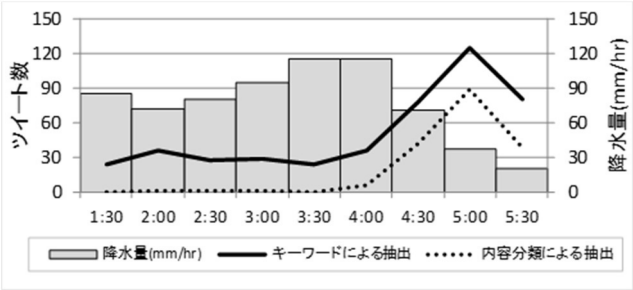
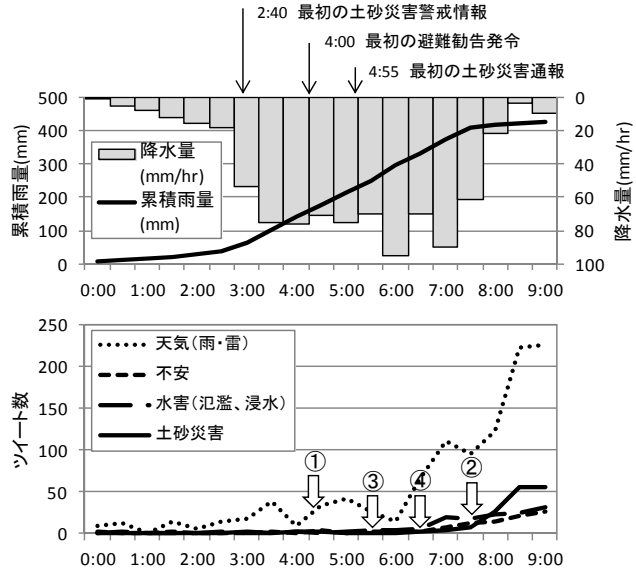


図 2-4　土砂災害に関連するツイート数の時系列変化（広島）

2.3.4　設定したキーワードによる状況把握

阿蘇市における災害の状況と，設定したキーワードで抽出したカテゴリーごとのツイート数の変化及び状況を直接的に示す内容のツイート例を図 2-5 に示す。



【状況を表すツイートの例（一部抜粋）】
①4:09　**雨**・**雷**半端ない!!**地響き**すごい
②7:17　阿蘇**やばい**らしい(°ｏ°;;　大丈夫か～？
③5:27　**床下**浸水いってそうです(´Д`)
④6:22　**土砂崩れ**で家が壊れた学校行けない。

図 2-5　ツイートによる災害情報の把握（九州北部豪雨）

天気カテゴリーのツイート数は継続して増加しており，激しい降雨が継続している状況が推察できる。また，7 時頃から水害や土砂災害カテゴリーのツイート数が急増しており，災害が発生又は危険が認識されている状況が推察できる。一方，激しい降雨や実際に土砂崩れが起きた状況等を具体的に示す内容のツイートも抽出された。これらの発言からはより具体的な現場の状況を把握できる。広島の結果を同様に図 2-6 に示す。阿蘇と同様，ツイート数の変化及びツイートの内容から，激しい降雨が続いている様子や災害が発生している状況が推察できる。

また，広島の場合は，阿蘇と異なり，土砂災害の発生より早い時間帯でツイート数の急増（例えば気象関連の 2 時 30 分頃や土砂災害関連の 2 時頃）が確認できる。要因として，人口規模が大きい都市部であり，分析対象となるツイート数が多いことが挙げられる。

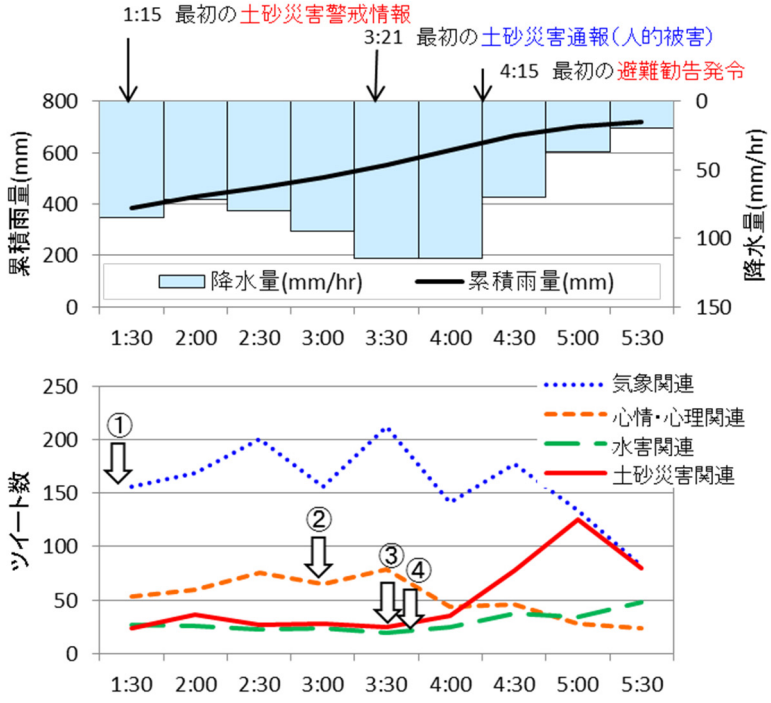


図 2-6　ツイートによる災害情報の把握（広島豪雨）

2.3.5　まとめ

本検討の結果，土砂災害に関連するツイート情報のカテゴリー分けおよびキーワード設定を行うことで，設定したキーワードから抽出したツイートの件数の変化や発言内容をもとに危険性の検知や災害状況の把握が可能であることが確認できた。

検討の対象外とされた位置推定ができていないツイートには「地鳴り」等の重要な語が含まれていることから，位置推定精度の向上による分析対象となる Twitter 情報の量及び正確性の確保，及び，設定したキーワードの組合せ（共起語等）により，災害に関連する発言内容をより的確に把握する技術や，その他強雨域の分布等，物理センサでの計測情報と重ね合わせて表示させることにより，防災情報としての有効性の向上等の検討が課題である。

2.4　警戒期から発災前における状況把握の可能性

前節までの調査結果から，Twitter 上では，発災に関する目撃情報や，警戒期において特徴的な情報が投稿されている場合があることがわかった。こうした情報を防災現場で有効に活用するためには，災害に関係するツイートを収集して可視化する一連の処理が，自動化されることが望ましい。一方，Twitter 上に投稿されるデータは多種多様であるため，防災に関連するツイートに絞り込む必要がある。また，災害に関連するデータに限ったとしても，災害規模が大きな場合においてはデータ量が増加する場合もあると想定される。このため，統計解析手法等を用いて，システムが自動的に情報を選別し，状況を可視化するような仕組みが必要だと考えられる。

そこで，本研究では，統計的手法を用いて，Twitter データから災害が発生する前の状況を把握できるかどうか調査した。具体的な調査項目は以下のとおりである。

- 状況が把握できる可能性を定量的に評価するため，特定の災害事例を対象として，Twitter 全量データに対して統計的な手法を適用し，警戒期から発災前に出現するデータの特徴を調査した。
- ソーシャルセンサとしての特性を整理するために，長期間のデータを用いて，雨に対する Twitter を介した人の反応を調査した。
- 以上の調査結果を踏まえ，統計的な手法の適用を行う場合の課題を整理した。

2.4.1　警戒期における Twitter データの定量的調査

(1) 調査の目的・アプローチ

本調査では，土砂災害が発生する前の警戒的な情報を Twitter データから抽出することを目的として，土砂災害の発災前後の Twitter データの全体像を掴み，Twitter データから抽出できる可能性がある情報を整理した。この調査のため，過去の土砂災害事例を対象として，Twitter の全量データを用いた俯瞰的な分析を実施し，統計的な観点から警戒期の情報を整理するアプローチをとった。警戒期に利用される気象データやカメラ画像等の多種多様な情報に加え，大量に存在する Twitter データから前兆現象等の情報を収集し，迅速に防災対応に役立てるには，将来的に処理を自動化することが望ましい。そのため，大量に存在する Twitter データからシステムが自動的に防災上有用な情報を抽出することを念頭においたものである。俯瞰的な分析としては，ツイート本文に対して話題分析を行って警戒期の話題の傾向を押さえた後，特定話題の傾向を分析するために共起語の調査を実施した。

(2) データ準備

本調査では，以下に示す 2 つの災害事例を対象として，Twitter データの分析を行った。

事例(a) 2012 年 7 月に発生した九州北部豪雨災害における熊本県で発生した土砂災害，2012 年 7 月 11 日 16 時から 12 日 12 時までの Twitter データを調査

事例(b) 2014 年 8 月に発生した広島市における土砂災害，2014 年 8 月 20 日 1 時から 5 時 30 分までの Twitter データを調査

Twitter データは，Twitter データを販売する業者から購入したもので，当該期間に日本でツイートされている全データを対象としたものである。本データに対してリツイートを除き，GPS，プロフィールおよびツイート本文の情報を元に都道府県を推定し，データの母集団とした。

(3) 話題分析

土砂災害の発災前後における，Twitter 上の話題を俯瞰し，出現する話題の経時的変化を調査した。具体的には，都道府県および時間帯毎にツイート本文中に出現する単語のランキングを作成し，話題分析を行った。ランキングの作成にあたっては，特定の場所と時間帯において特徴的に出現する単語を捉えるため，TF-IDF によるスコアを用いた。

TF-IDF は，単純に単語の出現回数の大小でスコアリングするのではなく，文書全体で出現しやすい単語のスコアを小さくし，特定の文書に集中して出現する単語のスコアを大きくする計算手法である。本調査では，全国のツイートを母集団とし，都道府県単位に仕分けた上で 1 時間毎にデータを集約し，TF-IDF の計算における文書の単位とした。この方法により，全データと相対的に比較した上で，特定の都道府県および時間帯に集中して出現した単語の傾向を可視化することができる。このため，災害が発生した都道府県において，災害が発生する前に限って特徴的に出現するキーワードを調査することができ，後の調査の土台となる情報を得られると考えた。

事例(a),(b)の調査結果を図 2-7 へ示す。調査の結果，災害発生前には，災害に関連するキーワードとして雷，停電，豪雨等のキーワードが頻出する傾向があることがわかった。

事例(a)H24熊本県土砂災害　↓4時頃より災害発生

	2012-07-11	2012-07-12						
	23:00-24:00	00:00-01:00	01:00-02:00	02:00-03:00	03:00-04:00	04:00-05:00	05:00-06:00	06:00-07:00
1	ハニラジ	雷	雷	雷	生活用品	添加剤	ビーバーナビーバー	苛苛
2	絵描き枠	一瞬停電	ジャミルさま	マモ	れいくん	バァンバリン	雷	休校
3	雷	味噌汁	光る	カミナリ	雑貨	インテリア	生活用品	白川
4	おやすみなさい	ポスト数	停電	果報	雷	生活用品	蛮族軍兵力報告	避難勧告
5	熊クマPイケメン	のむ	ピカピカ	停電	インテリア	雑貨	マムージャ	一読
6	ブサイカーズ	カミナリ	ゴルゴ	現れ	なつちゃん	是非一度	トロール	雷
7	恭一郎	光る	カミナリ	グーグルマップ	政木屋	エンジン音	雑貨	白川氾濫
8	寫字時的筆跡	夏菜子	右京さん	丼ぶり	名代	やみつき	カミナリ	リピーター様
9	マムージャ	停電	杏子ちゃん	光る	そば屋さん	避難勧告	インテリア	評価一覧
10	蛮族軍兵力報告	ぴーむ	おやすみなさい	ピカピカ	グーグルマップ	カミナリ	死者	氾濫
11	トロール	大雨洪水警報	マムージャ	城島	丼ぶり	詳細	救助	渡鹿
12	人生放棄	雷雨	蛮族軍兵力報告	雷雨	マムージャ	避難	雷雨	避難
13	特性	ピカピカ	生活用品	稲光	蛮族軍兵力報告	絶品	休校	学校休み
14	レろポッチャマ	おやすみなさい	トロール	落雷	トロール	燃費	詳細	自宅待機
15	ロボ製作所	ギュ	ギュ	東そば	ヒツジ	氾濫	鳴呼	安
16	通信欄	一瞬電気	雑貨	三宝	死者	雷	ファッション	数多い
17	Androidアプリ	ドオーモ	インテリア	ワンコイン	東そば	ファッション	豪雨	休講
18	死者	風雷益	雷雨	なおかつ	カレイド	アクセス	阿蘇	不調
19	マモル	絵描き枠	千早	地響き	エンジン音	工	メタル	水位
20	光る	ピカッ	玉ちゃん	敬う	食器	特徴	その他	雷雨

事例(b)H26広島市土砂災害　↓3時台に災害発生

	2014-08-20				
	01:00-02:00	02:00-03:00	03:00-04:00	04:00-05:00	05:00-05:30
1	何処	停電	午	ばなそにっく	せんぱぁ
2	譲る	雷	停電	避難	避難勧告
3	チケット	せんぱぁ	雷	避難勧告	ラグビー
4	停電	シコシコ	ファンネルミサイル	停電	景ちゃん
5	雷	雷雨	中華	占領	発令
6	ピカピカ	ゾロゾロ	落雷	のびる	わしわし
7	警報	ピカピカ	光る	氾濫	開設
8	シャロ	光る	せんぱぁ	消防車	雷雨
9	ノヤ	豪雨	あとアイチ	せんぱぁ	氾濫
10	光る	まき	けんじる	京葉線	新
11	ずばり	一印象	バグ	避難場所	マツピオ
12	入浴時間	落雷	冠水	四時	避難場所
13	願う	自発	サイレン	水位	根谷川
14	やや	殺生丸様	鳴る	安佐南区	パラダイステレビHD
15	ゴロゴロ	雛子	避難	パトカー	可部小学校
16	豪雨	雷鳴	きいちごパパ	vahrehvahさん	テレビ王国
17	一印象	大雨	竜巻注意報	幼児語	大林小学校
18	ビートルズ	ゴロゴロ	竜巻	ぴおか	微笑
19	扱い	信号	カミナリ	一次体制	可部学区
20	雨雲	カミナリ	ピカピカ	根谷川	大林学区

図 2-7 TF-IDF スコアのランキング（話題分析）

事例（a）熊本県と推定されたツイートで集中して出現した単語

事例（b）広島県と推定されたツイートで集中して出現した単語

事例(a)を対象に，武田ら（2013），武田ら（2014）で Twitter データの検索に用いたキーワード群と，土砂災害の前兆現象である土の臭いや地鳴りに関連するキーワード [16)17)]，および，前節での調査結果を元に選定したキーワードを用いて，それぞれのキーワードを含むツ

イート数の推移を調査した（図 2-8）。この調査結果から，量的な面では「怖い，やばい，すごい」「音」「停電」「注意報・警報」「強い雨の観察」に関する発言が発災前に増加することがわかった。また，ツイート数の増加率に関する調査結果から，土砂災害に関する前兆的ワード(小石，土臭い，地鳴り，地響き)が増加する時間帯があることがわかった。

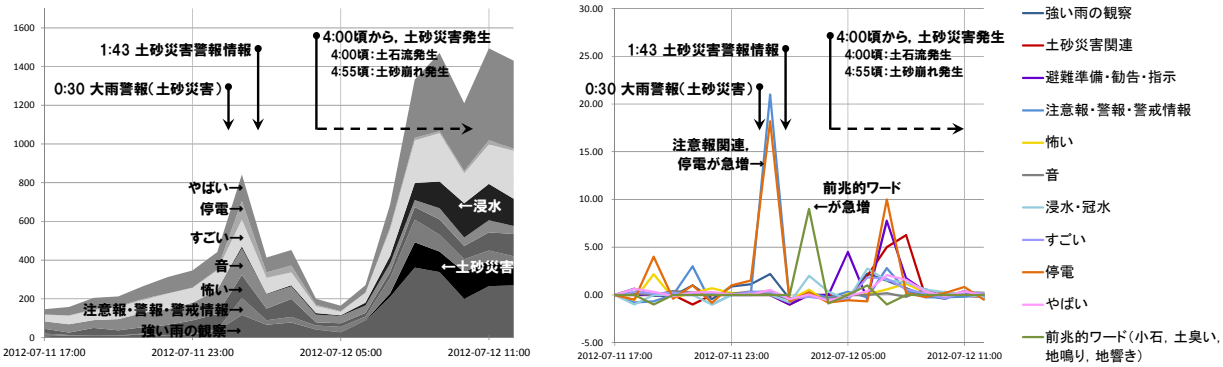


図 2-8 事例(a) 熊本県内のツイート数の推移

(左)ツイート数の推移，(右)増加率の推移　※1 時間単位で集計

(4) 雨発言に対する共起語の調査

前節の調査結果の中で，「怖い，やばい，すごい」「強い雨の観察」「音」に関する発言が警戒期に増加することに着目し，ツイートに含まれる共起語の推移を調査した。事例(a)に対する，検索キーワード別の共起語ランキングの推移を，図 2-9 へ示す。調査の結果，発災の直前から，災害に関する単語やネガティブな単語が，共起語の上位にランクインすることがわかった。また，全般的に「雨」や「雷」が共起語として頻出する傾向にあり，前節の調査と合わせて Twitter 上では「強い雨の観察」に関する情報を抽出できる可能性があると考察する。降雨の情報については，雨量レーダの情報を用いると正確な数値情報を入手できるが，Twitter 上の「強い雨の観察」は，住民による観察において普段とは違う，もしくは近年の経験から見て相対的に強弱を判断した上での発言と想定できる。より定性的な観点では，夜中に目を覚まして雨の恐怖をツイートするような状況であれば，十分警戒すべき状況であると考えることができる。

事例(a) H24熊本県土砂災害　　↓4時頃より災害発生

		2012-07-11		2012-07-12							
		22:00-23:00	23:00-24:00	00:00-01:00	01:00-02:00	02:00-03:00	03:00-04:00	04:00-05:00	05:00-06:00	06:00-07:00	07:00-08:00
雨	1	降る	降る	雷	雷	雷	雷	降る	雷	降る	降る
	2	くる	くる	降る	すごい	すごい	降る	猛烈	凄い	すごい	ない
	3	明日	明日	すごい	明日	寝る	すごい	みる	すごい	雷	凄い
	4	今日	すごい	くる	風	怖い	猛烈	解析	降る	凄い	今日
	5	ない	雷	寝る	寝る	降る	寝る	気象庁	ひどい	ひどい	雷
怖い	1	私	雷	雷	雷	雷	すぎる	覚める	雷	雷	雨
	2	笑	女子	雷	雷	くる	雷	知る	雷	雨	ない
	3	すぎる	ない	寝る	光る	雨	眠れる	話	今	やばい	家
	4	一番	笑	私	くる	寝る	寝る	目	せる	凄い	雷
	5	そう	思う	すぎる	寝る	すぎる	笑	くる	続く	すぎる	出る
すごい	1	思う	雨	雷	雨	雷	雨	雨	雨	雨	雨
	2	いい	ない	雨	雷	雨	雷	みたい	雨	雨	雨
	3	ない	思う	雨	ない	雷	雷	みる	雷	雷	今日
	4	やる	言う	今	雨	雨	雨	雨	今日	今日	お願い
	5	気	可愛い	いい	思う	寝る	今	今	気	雷	致す
やばい	1	笑	笑	雷	雷	雷	雨	雨	雨	雨	雨
	2	すぎる	行く	雨	雨	雨	すぎる	すぎる	雷	白川	笑
	3	そう	ない	すぎる	笑	寝る	これ	明日	これ	笑	白川
	4	くる	雨	笑	寝る	すぎる	寝る	ない	絶対	そう	行く
	5	いい	いう	死ぬ	すぎる	怖い	ノブ	氾濫	うち	家	学校
音	1	やる	気	雷	雷	雷	雷	雨	雷	雨	覚める
	2	感じる	聞こえる	落ちる	雨	すごい	寝る	眠れる	雨	笑	目
	3	時	よう	怖い	聞こえる	雨	雨	降る	大雨	そう	雨
	4	太平洋不知火楽団	立てる	くる	うるさい	光	あまり	思う	降る	思う	カミナリ
	5	思う	すごい	いい	考える	寝る	聞こえる	起きる	聞こえる	聞く	やる

事例(b) H26広島市土砂災害　　↓3時台に災害発生

		2014-08-20				
		01:00-02:00	02:00-03:00	03:00-04:00	04:00-05:00	05:00-05:30
雨	1	雷	雷	雷	猛烈	降る
	2	降る	降る	降る	降る	氾濫
	3	やばい	寝る	寝る	気象庁	猛烈
	4	音	くる	くる	影響	中心
	5	すごい	ない	やばい	土砂災害	局地的
怖い	1	雷	雷	雷	すぎる	ない
	2	すぎる	寝る	寝る	そう	考える
	3	そう	すぎる	笑	出る	言う
	4	見る	ない	そう	見る	そう
	5	雨	思う	ない	行く	俺
すごい	1	雷	雷	雷	雷	くる
	2	くる	雷	雷	くる	雨
	3	雨	思う	思う	雨	値する
	4	思う	ない	寝る	見る	コンタクト
	5	すぎる	私	雨	雷	朝焼け
やばい	1	雷	雷	雷	すぎる	雨
	2	すぎる	すぎる	寝る	寝る	笑
	3	笑	笑	すぎる	雨	大丈夫
	4	寝る	寝る	笑	家	避難勧告
	5	雨	そう	雨	笑	とる
音	1	雷	雷	雷	聞こえる	今
	2	雨	光る	目	雷	変
	3	光る	聞こえる	落ちる	光る	
	4	聞こえる	落ちる	光る	消防車	
	5	寝る	雨	覚める	寝る	

ネガティブかつ程度が大きいことを示す単語　　気象または災害に関連する単語

図 2-9　キーワード別共起語ランキング

(5) まとめ

本調査では，過去の災害事例を中心に俯瞰的なデータ分析を実施し，Twitter 上で警戒期から発災前に特徴的に出現する情報を整理した。調査結果から，Twitter 上では警戒期に特徴的な情報が出現する可能性があることがわかった。

2.4.2　雨発言を利用した潜在的ユーザー分布の可視化

(1) 調査の目的・アプローチ

Twitter データをソーシャルセンサと仮定して防災用途に利活用することを考えた場合，適用範囲を見極めるため，センサとしての網羅性や感度を把握しておく必要がある。具体的には，気象や災害等の事象の発生に対して，Twitter を介して人がどのような反応を示すのか，定量的な評価を行う必要がある。一方，前節までの調査で，Twitter 上では，雨に関するツイートが比較的多く存在することがわかった。このため，全国の雨に関するツイート数と雨量との関係を統計的に整理することで，全国の潜在的な Twitter ユーザーの分布や感度を可視化できるのではないかという，仮説を持った。この仮説を検証するため，著者が収集した 2012 年 7 月から 2015 年 6 月までのツイートを対象として，市町村毎の雨量とツイート数の関係を，統計解析手法を用いて考察した。

(2) データ準備

前処理として，Twitter データと雨量データをそれぞれ以下のように加工・集約した上で，場所（市町村単位）と日付をキーとして，データ結合を行った。総データ件数は，1,791,878 件となった。

1.　Twitter データの準備

「雨」を検索キーワードとして Search API を用いて収集した，2012 年 7 月 4 日から 2015 年 6 月 30 日のデータを母集団とした。母集団に対して，(1)GPS，プロフィールおよ

びツイート本文を用いて，ツイートの場所を市町村単位で推定し，(2)リツイート，伝聞やニュース等の目撃情報以外のツイートをテキスト解析処理・機械学習処理を用いて排除した。このデータ群に対して，市町村別に 1 日単位で集計し，集計テーブルを作成した。

2. 雨量データの準備

気象庁が提供する解析雨量データを雨量データとして用いた。前処理として，緯度経度情報から住所情報を推定する逆ジオコーディング処理を用いて，全てのメッシュを市町村へ対応付けた[5]。その上で，市町村毎に，領域内に含まれる雨域（雨量が 0 でないメッシュ）の平均雨量を当該市町村における平均雨量（時間雨量）として算出し，更に，平均雨量の 24 時間分の合計値を，当該市町村の 1 日雨量として算出した（図 2-10）。集約することで，雨の降り方についての詳細な情報は消失するが，Twitter データとの突き合せを考慮し，ある程度まとまった単位での集計とした。

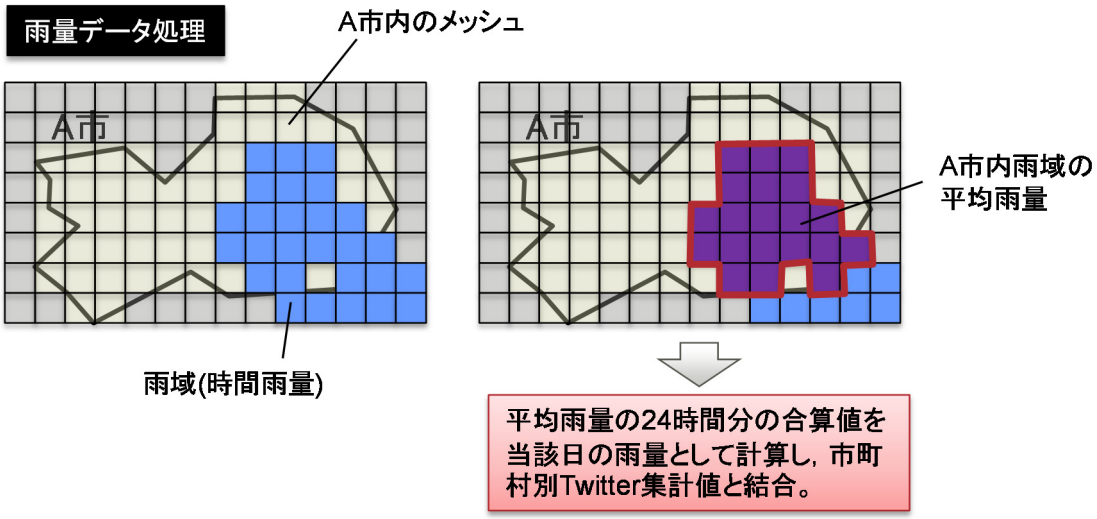


図 2-10　雨量データの準備方法

[5] この際，海上のメッシュは排除する処理を行った。住所起点の情報からの距離を用いた処理を行ったため，ある程度のノイズを含む。

(3) データ予備調査

雨に関するツイート数を変動させる要因として，①雨量の大きさ，②地域毎の人口，③雨量・人口以外の地域差の 3 つが関係するのではないかとの仮説を持った。この仮説を確かめるために，データを可視化し関係性を確認した。データの可視化においては，以下 2 つの観点から調査した。

1. 「①雨量の大きさ」の影響を確かめるために，特定の地域に絞って②および③の要因を排除して，雨量とツイート数の関係を可視化し傾向を調査した。ここで，雨量の大きさとツイート数に明確な関係がみられる場合，ソーシャルセンサとしての Twitter データが，雨量の大きさに対して反応が変わるものと想定される。
2. 「②地域毎の人口」による影響を確かめるために，市町村毎の人口差を是正したデータを用いて，雨量とツイート数の関係を可視化し傾向を調査した。ここで，1. で考察した結果と同様の傾向が見られた場合，地域毎の反応の違いが「②地域毎の人口」，すなわちソーシャルセンサの数のみの影響を受けると推測される。一方，1. で考察した結果と異なる傾向が見られた場合，「③雨量・人口以外の地域差」が存在するものと想定される。

「①雨量の大きさ」を確かめるために，鹿児島市のデータを対象として調査した。鹿児島市の全データの散布図を図 2-11 に示す。図より，雨量が大きくなると，雨に関するツイート数が増加する傾向にあると想定される。

次に，「②地域毎の人口」の影響を分析するための調査を実施した。用意した全データ（全国市町村のデータ）に対する散布図を図 2-12 に示す。図（左）は，1 日雨量とツイート数の関係を示したものである。一方，図（右）は，市町村毎の人口差を是正するために，ツイート数を市町村の人口で除して調整したものである（以下，発言率という）。本調査では，各集計値の 1 日単位で集計しているため，動的な昼夜の人口変動を考慮する必要はないが，人

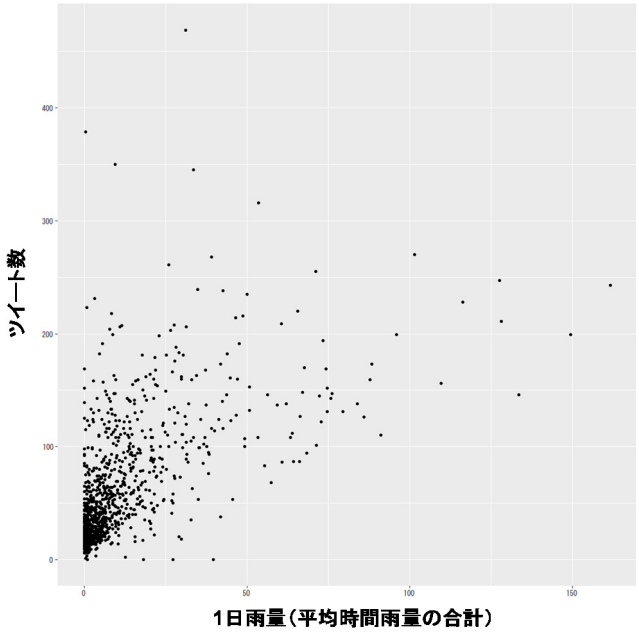


図 2-11　鹿児島市における集計データの散布図

ロのベースラインの算出方法を定めておく必要がある。そこで，総務省より公開されている統計情報を用いて，昼間人口と夜間人口の平均値を算出し，当該市町村の人口のベースラインとした[6]。

この図 2-12（右）をみると，雨量の大きさが大きくなったとしても，必ずしも発現率が増加するとは言えないことがわかった。また，雨量が小さい範囲においては，発現率のばらつきが大きい傾向にあった（図中の赤丸で囲った部分）。これらの傾向は，図 2-11 から考察された傾向とは異なるものである。したがって，雨に対するツイート数は，市町村人口や雨量以外の何らかの要因，すなわち「③雨量・人口以外の地域差」の影響を受けている可能性がある。これらの差異は，Twitter 上のデータを用いて統計処理を行う上で考慮すべき観点であり，定量的な評価が必要と考えられる。

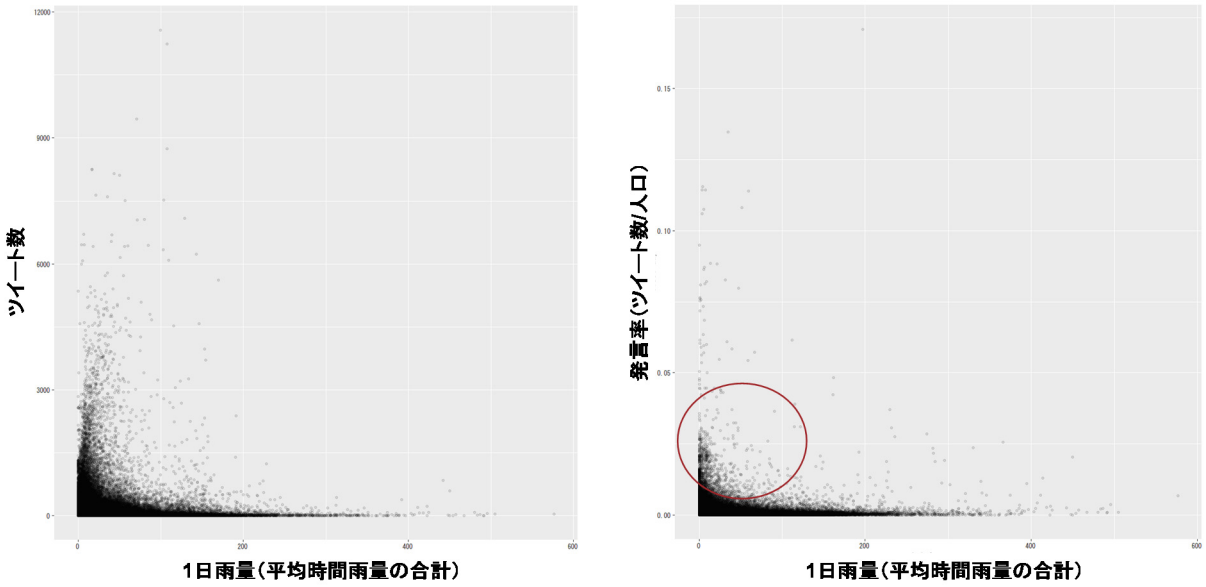


図 2-12　全データの散布図

（左）1 日雨量とツイート数の関係を示した図

（右）1 日雨量と人口調整後のツイート数（発言率）の関係を示した図

以上を踏まえ，定量的な評価を行うためには，市町村別の人口格差を排除した上で，統計モデルによる市町村の比較を行う必要がある。上の散布図では，ツイート数の人口調整のために，除算処理を実施した。しかしながら，被除数が小さく，かつ，人口の格差が大きい場合には，極端に人口が小さい地域において，調整後の値が過剰に評価されるという問題が発生する。この問題を，Small Number Problem といい，疫学等の分野において統計解析処理を

[6] これは，人の自宅外での活動時間帯のピークを朝 8 時～夜 8 時頃と想定したものであるが，あくまで仮説に過ぎない。全国市町村の人口変動を正確に把握することは現時点で困難であるため，本調査ではこの想定にしたがった。

行う際の留意点とされている[7]。この解消のための一つの方法として，階層的なベイズモデルを用いて，地域別の真の相対度数（人口調整後のツイート数）を潜在的なパラメータとして推定する手法が提案されている[8]。

本調査においても，階層的なベイズモデルを用いた推定処理を 1 日毎に実施し，全国市町村の人口調整後のツイート数を推定した上で，雨量データと突き合わせた。用意した全ての Twitter データを対象に,2 つの異なる方法で人口調整を行った場合の比較を図 2-13 に示す。図より，本調査で用いた集計データにおいても，単純な除算を用いて人口調整を行った場合，人口が小さい市町村では，集計値が 1-2 件となるような場合に，過剰に評価されていることがわかった。一方，階層的なベイズモデルを用いて推定した場合には，左記の問題は解消されているため，以後の解析では本データ処理を施したデータを用いた。

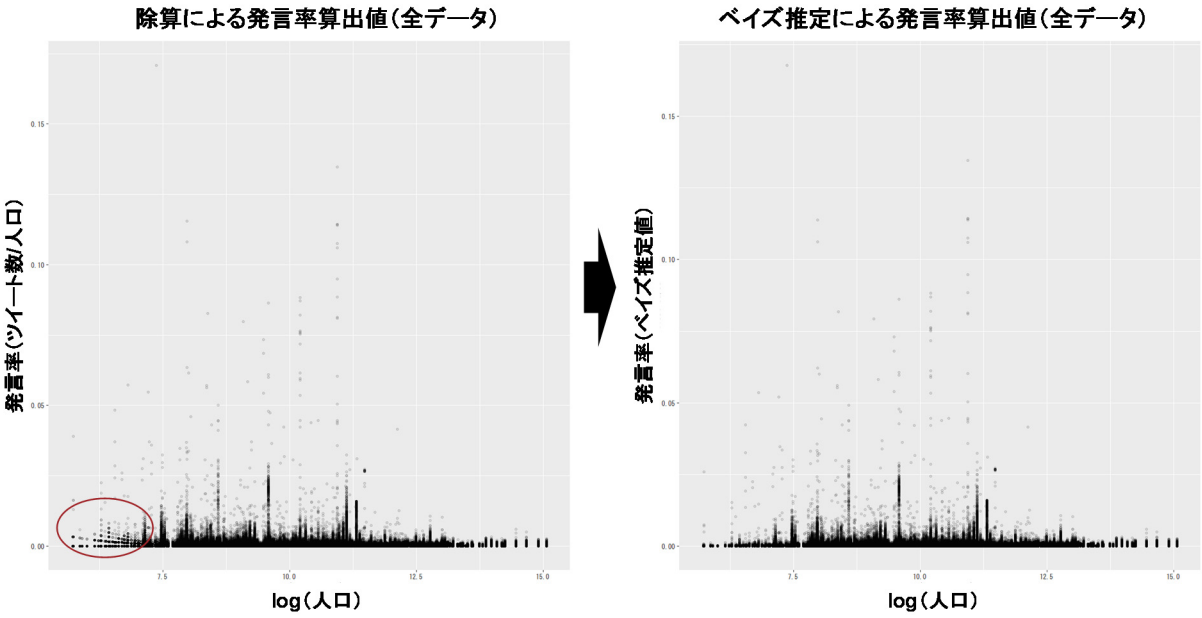


図 2-13　全ツイート数集計値における人口と人口調整後ツイート数（発言率）の関係

（左）除算によって人口調整した場合　※赤丸箇所で Small Number Problem が発生

（右）階層的なベイズモデルを用いた推定によって人口調整した場合

(4) 統計モデルを用いた解析

市町村別の雨に対する Twitter 上の反応をモデル化するにあたり，ツイート数を目的変数，雨量を説明変数とする回帰モデルの適用を検討した。統計モデルとして，以下 2 つの方法を比較した。

---

[7] 丹後ら（2007）[23]，瀬谷ら（2014）[24]を参照。

[8] 丹後ら（2011）[25]を参照。本調査においても，参考文献に記載されたモデル式を用いて推定した。モデルパラメータの推定には，Gelman et al.[26]を利用した。

モデル① 全国と市町村の 2 階層を仮定した階層的な線形回帰モデル（一般化階層線形モデル）を適用し，それぞれの市町村における全国共通パラメータからの差分を地域差とみなす方法。

モデル② 市町村それぞれに対して独立かつ非線形な回帰モデルを仮定し，独立に得られたパラメータを比較して地域差を推定する方法。

モデル①は，全国的に共通性を仮定してパラメータの交換可能性を考慮したモデルを想定することで，地域差そのものをパラメータとして推定することを狙ったアプローチである。このアプローチでは，全国の平均的なモデルをベースとして，地域差が傾きや切片のバラツキとして表現することができる。このため，市町村別の観測データ数が極端に偏る場合などにおいては，市町村別に独立的なモデルを仮定するよりも推定値が安定することが期待できると考えた。図 2-14 に，モデル①を用いて解析を行った例を示す。

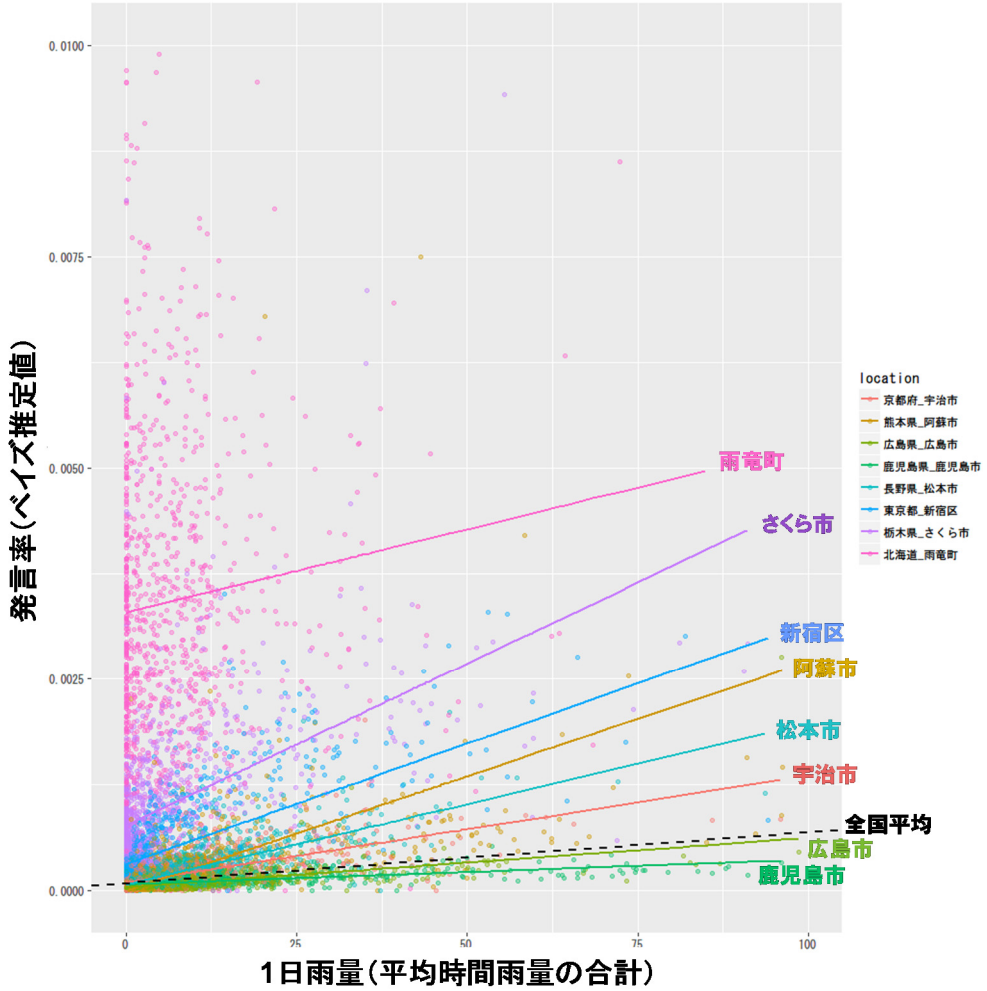


図 2-14 モデル①による解析例

（実線は市町村別，破線は全国レベルでの回帰モデル予測値を示す）

上の例では，線形モデルの傾き，切片ともにバラツキを持つモデルをあてはめた結果を示している。例えば，鹿児島市においては，全国平均と比較して傾きが小さく，Twitter を介した雨に対する反応が相対的に小さいものと考えることができる[9]。しかしながら，他の事例を含めて予備調査を行った結果，地域別に雨の降り方にもバラツキや偏りがあるため，すべての市町村で線形モデルを仮定するのは難しいことがわかった。

次に，モデル②の非線形な回帰モデルを用いた解析を検討した結果を示す。非線形な回帰モデルとしては，べき乗関数等の特定の関数を固定して考えるパラメトリックなモデルと，それらを仮定しないノンパラメトリックな手法がある。データ可視化等の予備調査の結果，市町村別に必ずしも同じような非線形な関係を持つとは限らないことが観察された。そこで，ノンパラメトリックな非線形モデルとして，市町村それぞれについて，独立な一般化加法モデル[10] を用いた。特定の市町村に対して，一般化加法モデルをあてはめた結果を図 2-15 に示す。

[9] 3 章に述べる現場実証実験におけるヒアリングにおいて，鹿児島県下では少々の雨では驚かないとのコメントを得ており，統計解析の結果とも一致する傾向となった。

[10] 一般化加法モデルは，それぞれの説明変数に対して非線形な基底関数を仮定して適用することで，非線形な統計モデルを構築する手法である。これにより，線形でなく平滑化されたモデルを構築することができる。平滑化には，平滑化スプライン関数やテンソル積が用いられる[27)]。

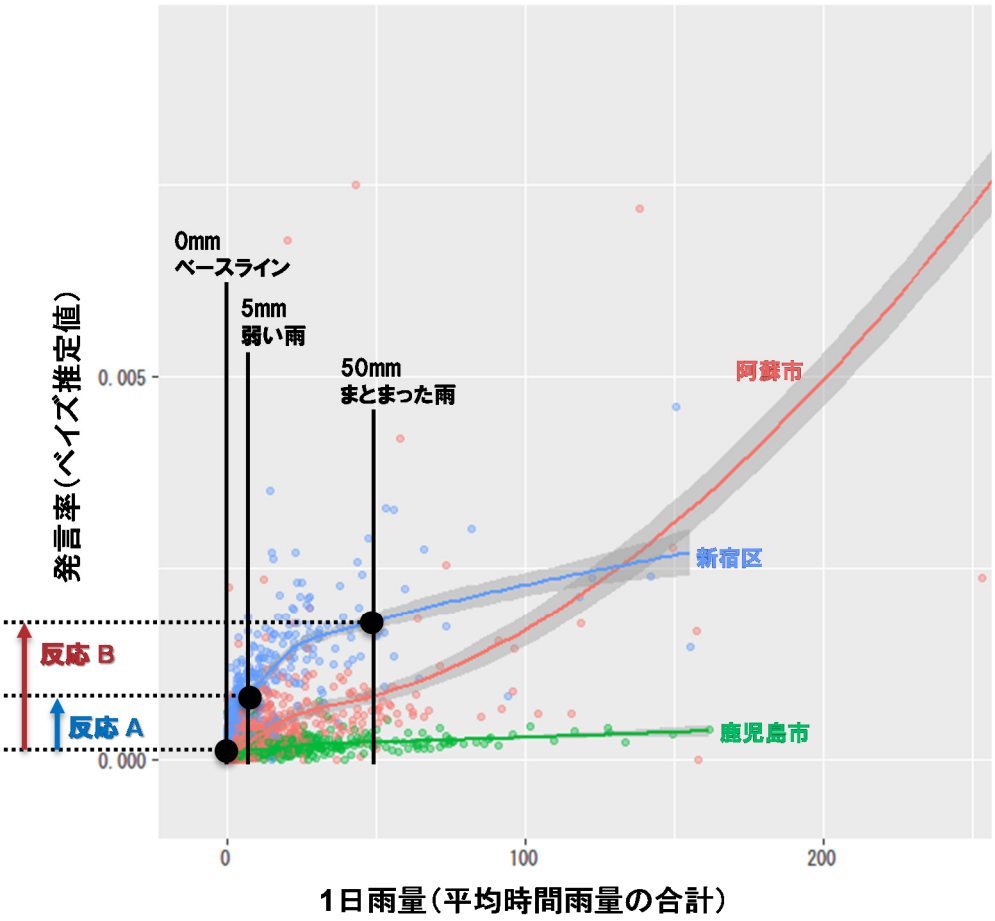


図 2-15　モデル②による解析例

（曲線は一般化加法モデルによる予測値，灰色の領域は信頼区間を示す）

推定結果より，1 日雨量とツイート数の関係として非線形な関係があり，かつ，市町村別に異なる関係性を持つと考察した。一方，市町村別に独立なノンパラメトリックなモデルを仮定すると，単純な方法では地域差を定量的に比較することができない。そこで，各市町村の推定モデルにおける，1 日雨量 0mm，5mm，50mm における人口調整後ツイート数のモデル予測値を使用して，1 日雨量 5mm と 0mm の予測値の差分を反応 A（弱い雨に対する反応），1 日雨量 50mm と 0mm の予測値の差分を反応 B（まとまった雨に対する反応）として算出する方法をとった。この方法を用いて，全国市町村の雨に対する反応の違いを数値化し，地図上に可視化したものを図 2-16 、図 2-17 に示す。図では，色が濃い程反応が大きいことを示している（等量区分にて 10 段階に色分けした）。

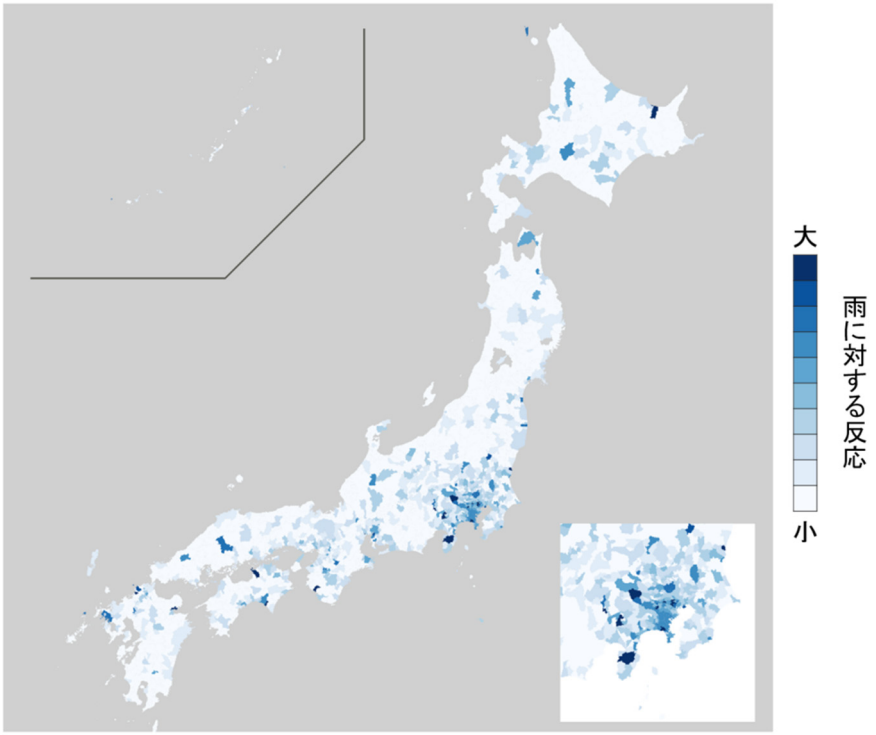


図 2-16　反応 A（弱い雨）に対する反応の分布

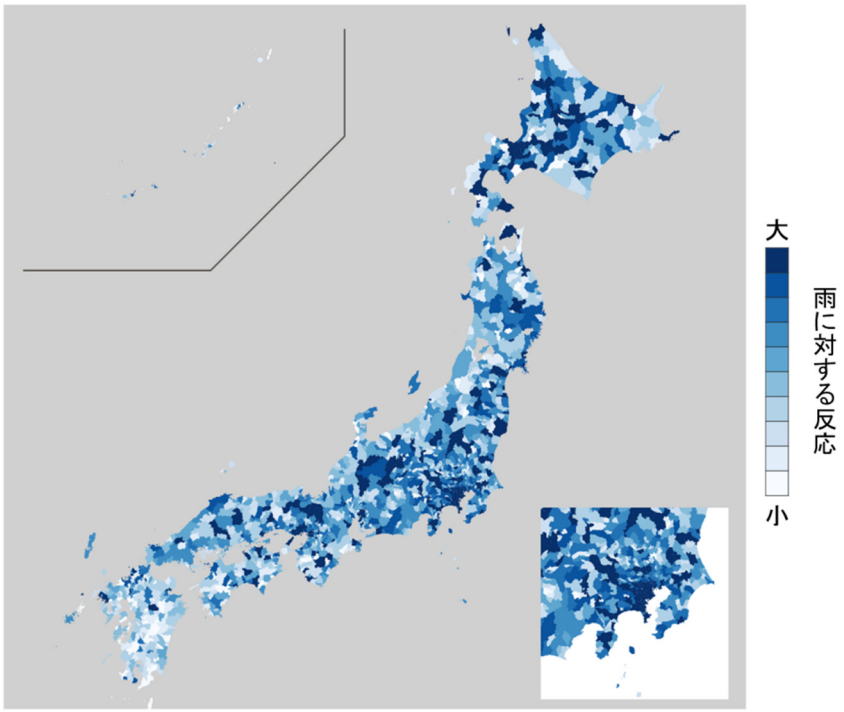


図 2-17　反応 B（まとまった雨）に対する反応の分布

図より，雨に対する反応は地域によって差異があることがわかった。九州南部や四国の一部地域では相対的に反応が小さいく，日常的に雨がよく降る地域では，雨に対して大きな関心を寄せない可能性がある。また，反応の大小はあるものの，全国的にみると，多くの地域で何らかの Twitter 上の反応があることが分かった。このため，統計的な前処理を行うこと

で，人口が小さい地域においても，武田ら（2013），武田ら（2014）で用いた統計解析手法[16)17)]を適用して，Twitter 上の反応を定量的な情報として抽出できる可能性がある。

なお，ここで示した地域差には，地域別の人の反応の差に加えて，テキストデータを用いて場所を推定する上での曖昧性を含んでいると想定される。本調査では，プロフィールやツイート本文に含まれる住所やランドマークの情報を用いて場所を推定した。この推定処理においては，単語と場所をマッチングさせる辞書の影響を受けることに加え，自然言語処理の特性から一部確率的な処理を含んでおり，出力結果に偏りが発生する可能性がある。したがって，本調査で得られた地域差に相当する解析結果は，純粋に人の反応の差のみを示すものとは言えないが，一連の処理プロセスの結果として捉えた場合に，システムの出力結果の偏りを示すものであると考察した。このため，統計的な処理を用いて，全国のツイート数を評価する場合においては，地域の人口差に加えて，本調査で得られた潜在的な反応の差（例えば，降雨に対する地域別の人の反応の差）を考慮する必要があると想定される。

(5) 本調査で得られた解析結果の活用例

過去の災害事例として，2014 年 8 月 20 日に発生した広島市の土砂災害事例を取り上げる。本事例では，2.4.1 節に示すように，土砂災害が発生する数時間前から，Twitter 上で強雨に関して言及するツイートが投稿される傾向にあった。そこで，「雨」を含むツイート群に対して，更に「強い」「怖い」「やばい」等の単語を共起語として含むツイートに絞り込み，2.4.1 節で示したノイズ排除処理および場所推定処理を行ったところ，広島県内で 01:00 から 02:00 の間に 65 件のツイートが存在することがわかった。

こうした傾向を，統計的な処理を用いて自動的に抽出できるとするならば，災害発生前の現場で起きている状況を，システムが自動的に検出してアラートを上げることができる可能性がある。ただし，実際の運用を考えた場合には，どの場所で日常と異なる状況が発生しているか不明であることから，全国のデータを横並びで把握する必要があると想定される。そこで，当該時間帯の全国集計値を比較したグラフを，図 2-18 へ示す。

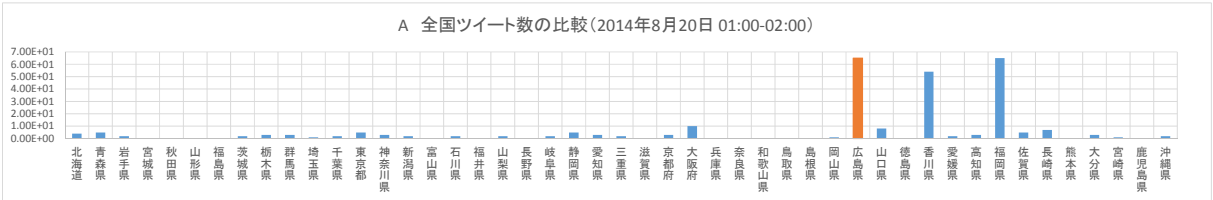


図 2-18 強雨観察ツイート数（2014 年 8 月 20 日 01:00-02:00）

上図より，確かに全国的に見ても広島県のツイート数が突出していることがわかる。一方，全国のツイート数を，相対的な観点で比較する場合には，人口による調整を実施する必要がある。そこで，本節で議論した方法を用いて，当該時間帯の全国ツイート数の人口格差を階層的なベイズモデルを用いて是正した結果を図 2-19（上）に，独立な一般化加法モデルを用いて地域の反応差を考慮して調整した結果を図 2-19（下）に示す。

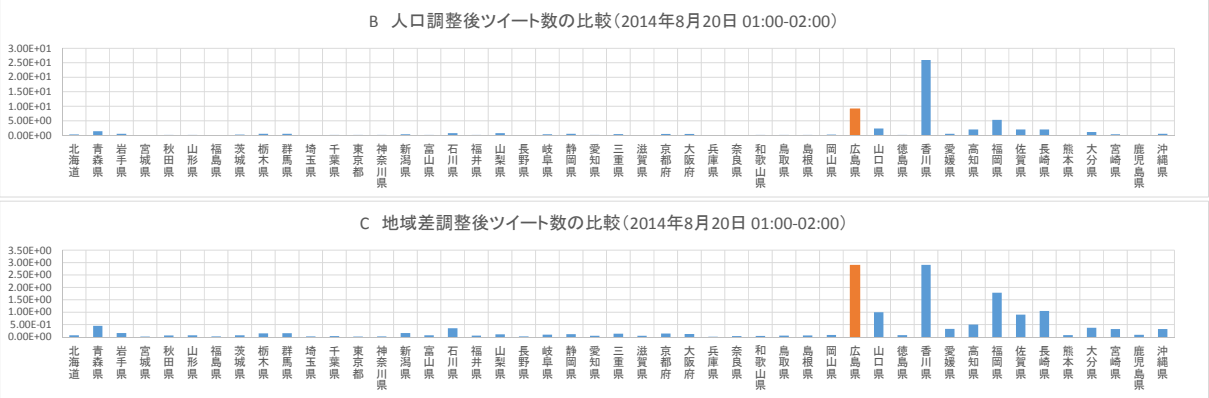


図 2-19　調整後の強雨観察ツイート数（2014 年 8 月 20 日 01:00-02:00）

調整後の強雨観察ツイート数を見ると，単純に人口調整のみを適用した場合には，広島県よりも香川県が，相対的にツイート数が多くなる結果となった。この結果に対して，地域差を考慮した調整を行うと，広島県，香川県がそれぞれ相対的に大きく評価され，統計処理を行った場合に，どちらの場所においても異常検知等の手法[11]を用いて情報を抽出できるものと想定される。なお，香川県におけるツイートは，当時，まんのう町やさぬき市，東かがわ市で 1mm 全後の雨量が観測されており，これに対する反応と考えられる。しかし，強雨が観測された広島県と比べ，雨量が小さかったにも関わらず同程度のツイート数となっている。この要因としては，場所の推定処理の誤りや偏りによって，広島県のツイートの一部が香川県と推定されたことや，伝聞や意見等の直接的な観察ではないノイズとなるツイートが存在していたことの影響が大きかった可能性が考えられる。

(6) まとめ

本調査では，統計モデルを用いて，雨という自然現象に対する Twitter を介した人の反応をモデル化し，ソーシャルセンサとしての Twitter の特性を調査した。また，全国市町村の潜在的な Twitter ユーザーの分布を可視化することを試みた。

この調査の結果，Twitter を介した人の反応は，人口以外の隠れた要因によって差異が生じる可能性があることがわかった。また，調査の過程において，統計的な観点から人口調整を行う場合には，Small Number Problem への考慮が必要なことがわかった。

これらの観点は，Twitter データをソーシャルセンサと見なして統計解析処理を行う場合において，考慮すべき事項であると考えた。具体的には，地域別に Twitter 上の話題（災害，気象等）の発生を自動的に抽出することを考える場合，人口格差と隠れた地域差を考慮することで，全国的に偏ることなく情報を抽出できると想定される。この仮説は，特定の事例において効果があることがわかった。また，Twitter 上では全国的に自然現象等に関するツイ

[11] データ群において，通常のデータとは発生量や振る舞いが異なるデータを検出することを異常検知という。統計的・確率的な手法が数多く提案されている [28),29),30)]。

ートの発生が期待できることがわかった。

以上の考察は，約 3 年間のデータを使った実験より導いたものである。しかしながら，気象現象の周期性や土砂災害の発生頻度を考慮すると，短い期間のデータに基づく調査結果である。降雨に対する Twitter ユーザーの反応は,雨期（6 月～10 月）とそれ以外の時期における違いや,強雨の程度による差異なども考えられる。また,隠れた地域差には,地域や Twitter ユーザーの防災意識などが影響することが考えられる。そのため，ソーシャルセンサの防災分野の活用を検討するにあたっては，社会実装を行いながらデータを収集し，今後も継続的な分析調査を行っていく必要があると考える。

2.5　土砂災害の前兆現象等情報の収集・把握手法の可能性と課題

2.5.1　Twitter データに対する統計的アプローチにおける課題

(1) アプローチ

本共同研究では，土砂災害警戒期の Twitter データを用いて，統計的な観点で情報を抽出するための基礎的な調査を実施し，以下の考察を得た。

① 土砂災害の発生前において，Twitter に投稿されるツイート文に，特徴的な表現（キーワード）が出現する可能性がある。【2.4.1 節より】
② Twitter データをソーシャルセンサと仮定した場合，センサの感度は人口以外の隠れた要因により変動する推定されるが，その要因を統計的にモデル化することで，センサ感度の全国的なバラツキを是正できる可能性がある。【2.4.2 節より】

上記①から，Twitter データを活用することにより，土砂災害発生前の特徴的な情報を抽出できる可能性があることが分かった。このため，Twitter 上の関連ツイートをソーシャルセンサと見立て，武田ら（2013），武田ら（2014）に示すような異常検知の手法[16)17)]を適用することにより，土砂災害が発生する前の警戒的な情報を自動的に抽出することが可能であると考える。この考えに基づいた，Twitter データ活用のアプローチを以下に示す。

- 上記①で明らかになった特徴的なキーワードを元に Twitter データの収集を常時行う。
- 収集したデータに対して，武田ら(2013)に示すノイズの除去処理[16)]および武田ら(2013)，武田ら（2014）に示す場所の推定処理[16)17)]を経て地域別に集計する。その後，上記②のセンサ感度のバラツキを是正する処理を実施し，異常検知処理における特徴量とする。
- 上記特徴量を入力として統計的な異常検知手法を適用し，平常時と異なる状態を自動的に検知し，システムの利用者等にアラートを送信する。異常検知手法としては，一例として，平常時のデータを用いてモデル化し，時間・地域(空間)的に突出したデータを捉える手法の適用が想定される[16)]。
- アラートが発生した場合，システムの利用者がデータの内容を把握できるようにするため，異常検知に寄与したツイートや関連する投稿写真，キーワード別の集計値の推移等をアラート情報と紐づけて蓄積する。

上記アプローチにより，土砂災害に関連のある Twitter データを常時収集しながら，統計的手法を用いて自動的に監視するシステムを構築することができると想定する。

(2) 課題

前項で示したアプローチを防災業務に適用するには，以下の課題を解決する必要がある。

1. 異常検知手法の継続的な検討

2.4.1 節の調査で明らかになった特徴的なキーワードの中には,雨などの自然現象に対する観察を意味するものが多く見受けられた。このように，人が自然現象を観察して目撃した際のつぶやきを利用する異常検知では，過去に経験したことがないような豪雨や災害に遭遇した場合の言葉の表現，防災意識等の変化による災害に関するツイート数や言葉の変化などが影響する可能性が考えられる。このため，異常検知手法のアルゴリズムを検討する上でその妥当性を担保するためには，より長期間のデータを用いて継続的な調査を行う必要があると考える。また，本研究で取り扱った土砂災害の発生間隔は，これまでソーシャルメディアが利用されてきた期間と比べて長いため，アルゴリズムの精度検証は継続的に実施されるべきである。

2. 場所推定精度・細かさの改善

Twitter 上に投稿されるツイートの多くは位置情報（GPS）が付与されていない場合が多く，ユーザーのプロフィールやツイート本文から投稿者の所在地を推定する必要があるため，リアルタイムで把握できる情報の細かさは現時点では市町村レベルの情報となっている [17)]。一方，ツイートから得られた警戒的な情報を防災情報として活用するためには，より詳細な場所の情報があるのが望ましい。このため，各ツイートもしくはツイート群から得られた情報の発生源をより細かく推定するための技術開発が必要である。この課題の解決に向けては，仮説として，Twitter データ以外の気象情報，土壌雨量指数など動的に変化するメッシュ情報を活用し，場所を絞り込むなどのアプローチが考えられる。

3. 適用条件の整理

武田ら（2015）の調査により，人口の少ない地域であっても，ある程度 Twitter の利用者が存在することがわかっている [17)]。しかしながら，人口が著しく少ない地域や，起きている人が相対的に少なくなる深夜時間帯などにおいては，災害の警戒期に外界の状況を観察する人の数が減少するため，ソーシャルセンサの網羅性や感度に影響すると想定される。

また，災害の外力や災害そのものの規模によって，ソーシャルセンサの感度も異なると考えられる。このため，多種多様な災害事例を通して，本アプローチを検証し，適用条件を整理することが課題である。適用条件の整理においては，異常検知手法の汎用的な性能を検証するために，多くの事例を用いて統計的な整理をするべきである。

上に示した課題事項は，主として Twitter データを軸とした災害情報抽出手法やその適用範囲に関する課題事項である。一方，他の防災情報との関連を考えた場合，以下のような課題がある。

4. 防災情報としての表現方法の検討

防災分野における新たな情報ソースとして Twitter データを位置付けた場合，他の防災情報とどのように関連付けて活用するか，統合的な情報管理の方法を検討する必要がある。この検討のためには，適用条件の整理結果を起点として，Twitter から得られた情報を効果的かつ効率的に防災関係者が状況を把握できるようにするような新たな表現方法を検討することが必要である。具体的には，Twitter から得られた情報が他の防災情報と並列で提示された場合に，防災関係者にその位置付けをわかりやすく，かつ，誤解のないように提示することを目的とするものである。特に，Twitter の情報は曖昧性を含むため，統計的手法を用いて情報の確からしさを数値的に表現したり，ツイート本文に含まれる一次情報をコンパクトに提示したりするなどの工夫が必要と想定する。

5. 統合的な防災データの解析手法の開発

長期的な観点として，Twitter データ，気象情報，土壌雨量指数等のデータを災害に係る統合的なデータ群に位置付け，新たなデータ解析手法を駆使し，一連の情報から土砂災害が発生する前の警戒的な情報をより精度よく抽出するための検討が考えられる。一方，情報の性質，観測時間の間隔のズレ，場所の細かさが異なる多種多様なデータを活用することになるため，本共同研究や関連研究で示した手法とは別に，新たな手法を開発する必要があると想定する。