経済・社会への分野横断的研究会
キヤノングローバル戦略研究所
2017年9月25日（月）～ 26日（火）

ツイッターデータを用いた
意識の違いによる人の行動パターン分析
(Comparison between Spatial Distributions of Tweet Base and Population in Japan)

藤本　祥二　（金沢学院大）
石川　温　（金沢学院大）
水野　貴之　（国立情報学研究所，CIGS）

イントロダクション

- 国勢調査や国民生活調査
  - 大規模な調査（日本でも5年に一度）
  - 頻度を上げたい、出来ればリアルタイムに
  - 発展途上国などでは調査が難しい（インフラ）
- 今回の研究
  - 日本の位置情報付きツイッター投稿データを用いて、人の流れを把握
  - 国勢調査や生活調査の結果と比較
  - 公的な統計調査をどの程度再現できるのか検討

目次

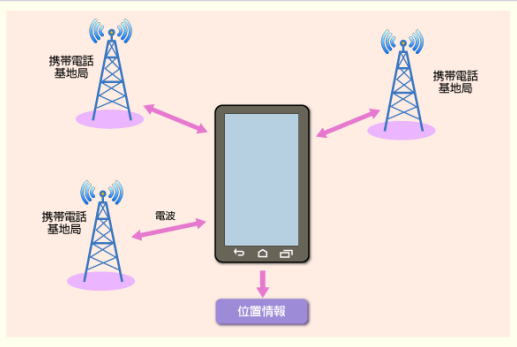
- イントロダクション
- データ
  - ツイッターデータ
  - 国民生活時間調査データ
- データ分析
  - 人口分布の再現
  - 起床在宅率の再現
- まとめと今後の課題

目次

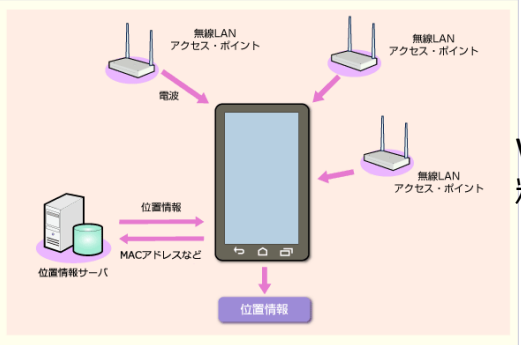
- イントロダクション
- データ
  - ツイッターデータ
  - 国民生活時間調査データ
- データ分析
  - 人口分布の再現
  - 起床在宅率の再現
- まとめと今後の課題

位置情報付きソーシャルデータ

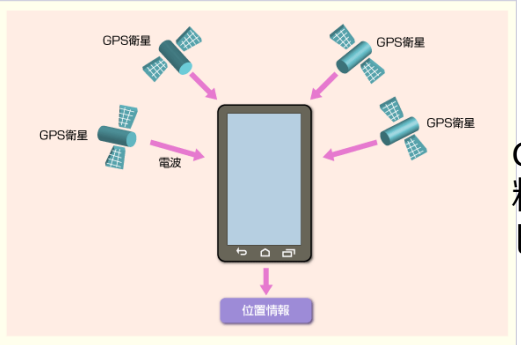
- 携帯端末の位置情報
  - 精度は右の通り
- 様々なアプリと連携
  - 地図アプリ
  - ソーシャルネットワークサービス（SNS）
  - 端末所持者の自由意志で位置情報の提供



携帯電話基地局
精度：数km



Wi-Fiアクセスポイント
精度：十数m～約200m



GPS
精度：数m
ビルの影や地下は×

画像元：第6回　迷わないためのAndroid端末の位置情報・GPS設定術
http://www.atmarkit.co.jp/ait/articles/1310/23/news120.html

使用ツイッターデータ

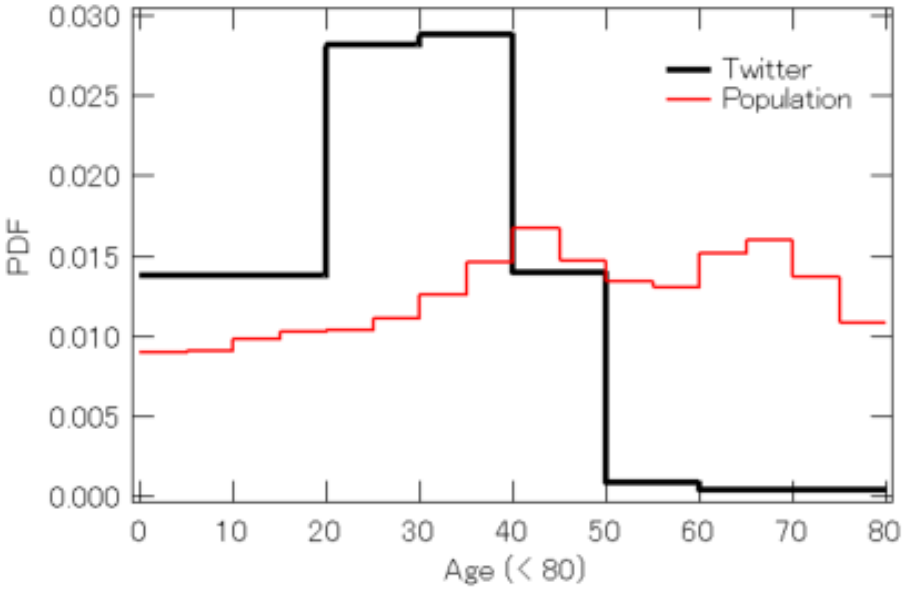
- 緯度・経度の位置情報が付いた
  ツイッターデータ
  （位置情報付きのものだけを入手）
  - ユーザID，年月日，時分秒，緯度，経度
    の項目を分析に用いる
  - 日本の領土を全て含んだ長方形の領域
    北緯：20.4～45.6　　東経：122.9～154.0
  - 期間
    2014/03/04～2015/04/26　　（419日間）

４０万ユーザーのプロファイル

位置情報付きユーザーの
約80%程度をカバー

ツイート内容からのプロファイリング情報

- ユーザーID，都道府県，スクリーンネーム，年齢（７分類），結婚，年収（８分類），性別，飲酒，喫煙，カメラ等３０種類の趣味の有無，業界（１６分類），職種（１０分類），在住地方（８分類），車移動等６種類の移動手段の有無，職業（１２分類），役職（９分類）



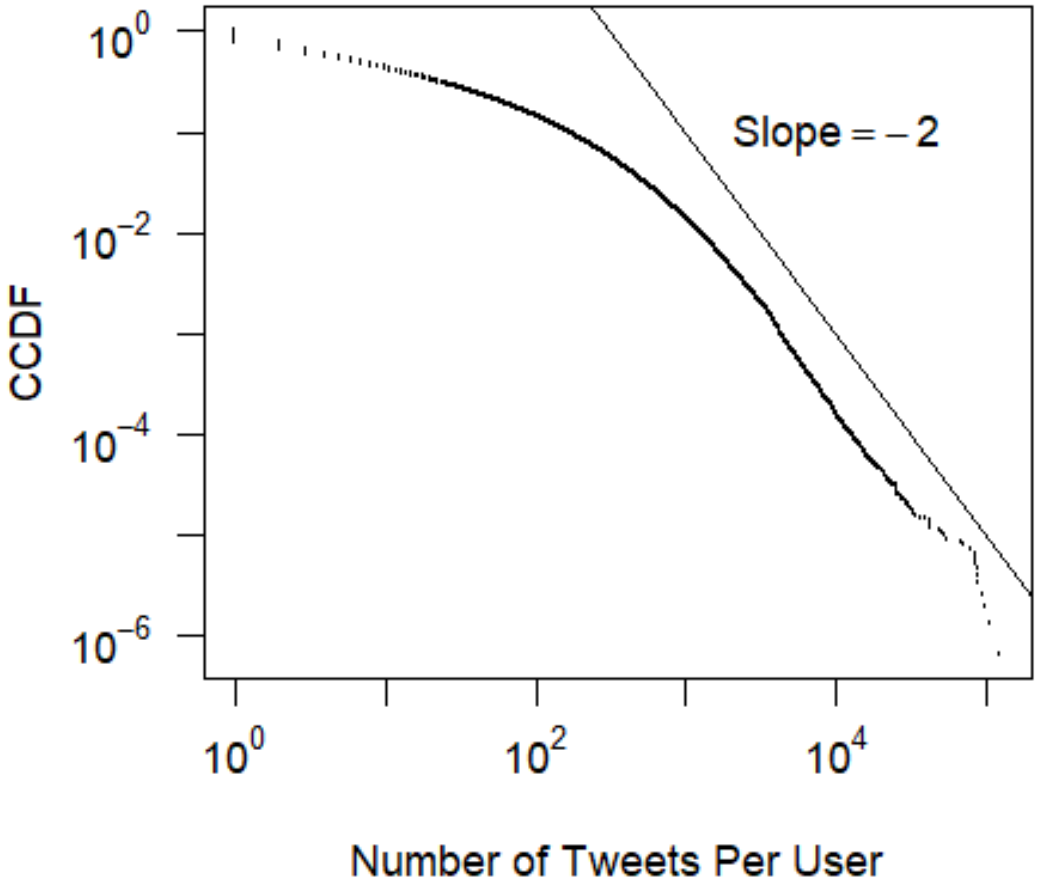
年齢＼性別	男	女	不明	合計
_19	51229	18651	11133	81013
20_29	39517	31503	11622	82642
30_39	57243	17755	9567	84565
40_49	31281	5878	3916	41075
50_59	1708	519	254	2481
60_	1551	222	218	1991
不明	60393	30916	14924	106233
合計	242922	105444	51634	400000

⟺ 20代30代が最も多い

ツイート数とユーザー数

全ツイート数：
127,577,069

全ユーザー数：
1,521,584



- 上位1%は全期間で1,000ツイート
  1日平均2.5回以上
- 下位90%は全期間で100ツイート
  4日に1回程度

月別アクティブユーザー数(MAU)

位置情報付きデータの、月別ツイート数と、MAU

年/月	ツイート数	MAU	年/月	ツイート数	MAU
2014/03	8,653,616	247,589	2014/10	8,103,683	268,674
2014/04	9,055,384	243,923	2014/11	7,445,254	268,025
2014/05	9,715,331	262,150	2014/12	10,571,060	363,151
2014/06	9,041,571	278,228	2015/01	9,607,290	346,343
2014/07	8,648,575	250,079	2015/02	8,924,144	377,909
2014/08	9,303,617	266,631	2015/03	11,745,895	446,305
2014/09	7,610,038	245,515	2015/04	9,151,611	382,600

MAU（Monthly Active User）：1か月に1回以上ツイートしたユーザーのこと

Twitter JP 公式発表MAU

- 2015/12　　3500万人
- 2016/11/01　4000万人突破

MAUの1～2%程度が位置情報を提供

目次

- イントロダクション
- データ
  - ツイッターデータ
  - 国民生活時間調査データ
- データ分析
  - 人口分布の再現
  - 起床在宅率の再現
- まとめと今後の課題

国民生活時間調査

- NHK放送文化研究所（世論調査部）
  - 1960年から，5年に1度の調査
  - 2015年10月13日(火)～26日(月)の調査
    - 住民基本台帳から層化無作為２段抽出
      12,600人（12人×150地点×７回）
    - サンプル数

	指定サンプル数	調査有効数	率
平日	18,000	11,056	61.4%
土曜	3,600	2,195	61.0%
日曜	3,600	2,170	60.3%

調査票2015年

10月13日（火曜日）2 1

午前の生活

◆ 13日（火曜日）のお天気は？
番号に○をつけてください

1．雨はふらなかった
2．一時、雨がふった
3．ほとんど１日中、雨がふった

◆ 13日（火曜日）のあなたの行動は？
起きた時刻・寝た時刻を記入し、以下の行動をしたかどうか○をつけてください

起きた時刻 →（午前 ・ 午後 時 分ごろ）
寝た時刻 →（午前 ・ 午後 時 分ごろ）
朝食→食べた，食べなかった／昼食→食べた，食べなかった／夕食→食べた，食べなかった
仕事 → した，しなかった
在宅状況 → 自宅から外に出なかった，自宅にいたり出かけたりした，まったく自宅にいなかった

起きた時刻からではなく
午前０時から記入してください

「自宅にいた時間」の欄も
忘れずに記入してください

01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48

午0前時 30 1時 30 2時 30 午3前時 30 4時 30 5時 30 午6前時 30 7時 30 8時 30 午9前時 30 10時 30 11時 30 12時

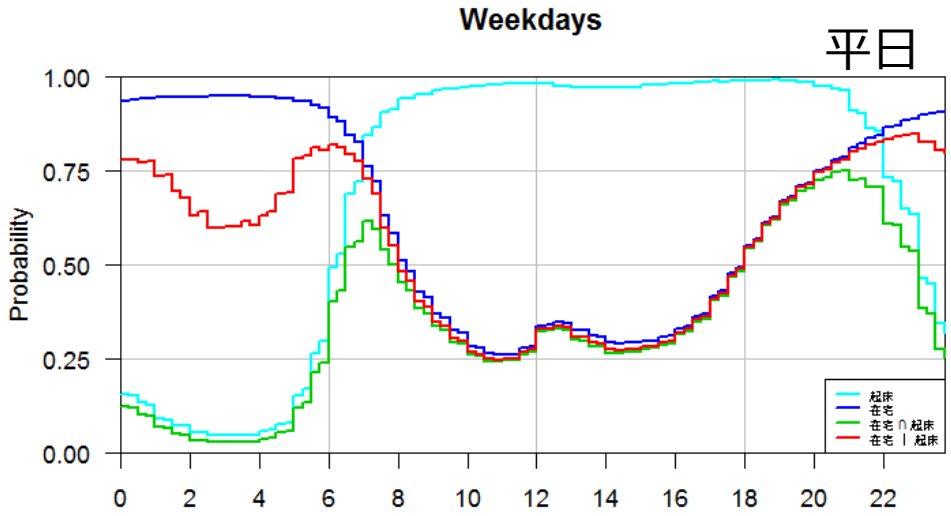
自宅にいた時間	01
すいみんをとる（30分以上）	02
洗面・入浴・着替えなどの身のまわりの用事	03
食事をする	04
通勤（往復）	05
仕事をする	06
仕事上のつきあい	07
通学（往復）	08
授業・学校の行事・部活動・クラブ活動	09
宿題・予習・復習・塾の勉強	10
炊事・掃除・洗濯をする	11
買い物をする	12
子どもの世話をする	13
その他の家事をする（内職・用事・病人の世話など）	14
社会参加（PTA・地域活動・冠婚葬祭・ボランティアなど）	15
つきあい・交際・おしゃべり	16
スポーツをする	17
行楽地へ行く・散策する・ぶらぶら歩く	18
趣味・娯楽・けいこごとをする・あそぶ	19
趣味・娯楽・けいこごと・あそびでインターネットを使う	20
テレビを見る	21
ラジオを聞く	22
新聞を読む	23
雑誌・マンガ・本を読む	24
CD・デジタルオーディオプレイヤー・テープを聞く	25
ビデオ・HDD・DVDを見る	26
休息をとる	27
療養する・診療を受ける	28
その他（分類できない行動は、すべてここに具体的に書いてください）	29

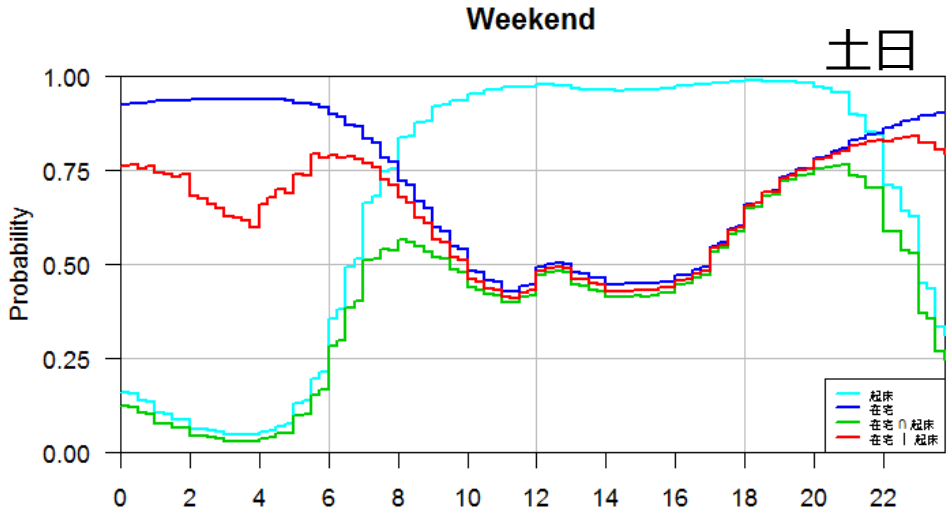
15分単位で塗りつぶす

午0前時 30 1時 30 2時 30 午3前時 30 4時 30 5時 30 午6前時 30 7時 30 8時 30 午9前時 30 10時 30 11時 30 12時

01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48

国民生活時間調査の比較対象項目





水色：*P*(起床)
　　1－*P*(睡眠)

青色：*P*(在宅)

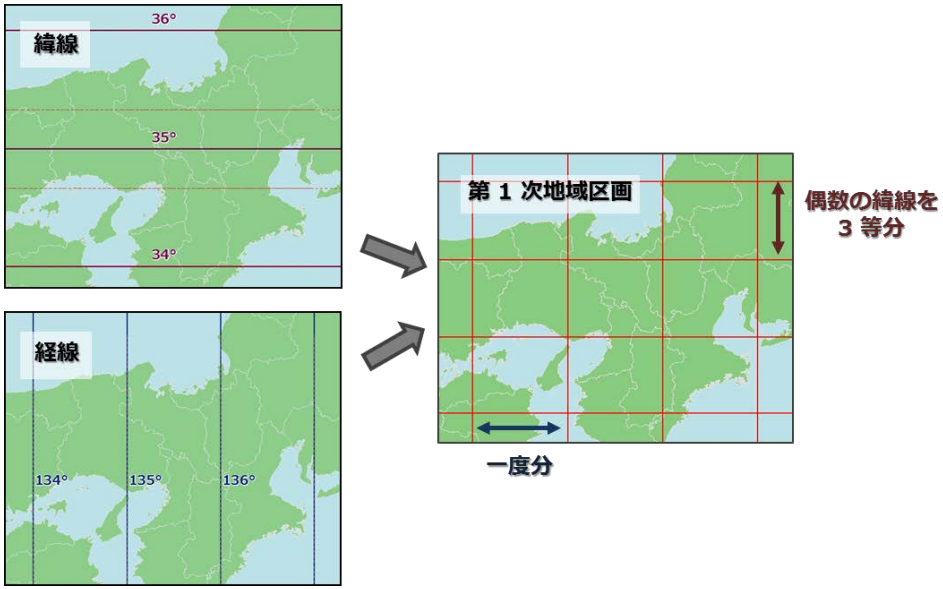
緑色：*P*(在宅 ∩ 起床)

赤色：*P*(在宅｜起床)
　　＝緑色／水色

目次

- イントロダクション
- データ
  - ツイッターデータ
  - 国民生活時間調査データ
- データ分析
  - 人口分布の再現
  - 起床在宅率の再現
- まとめと今後の課題

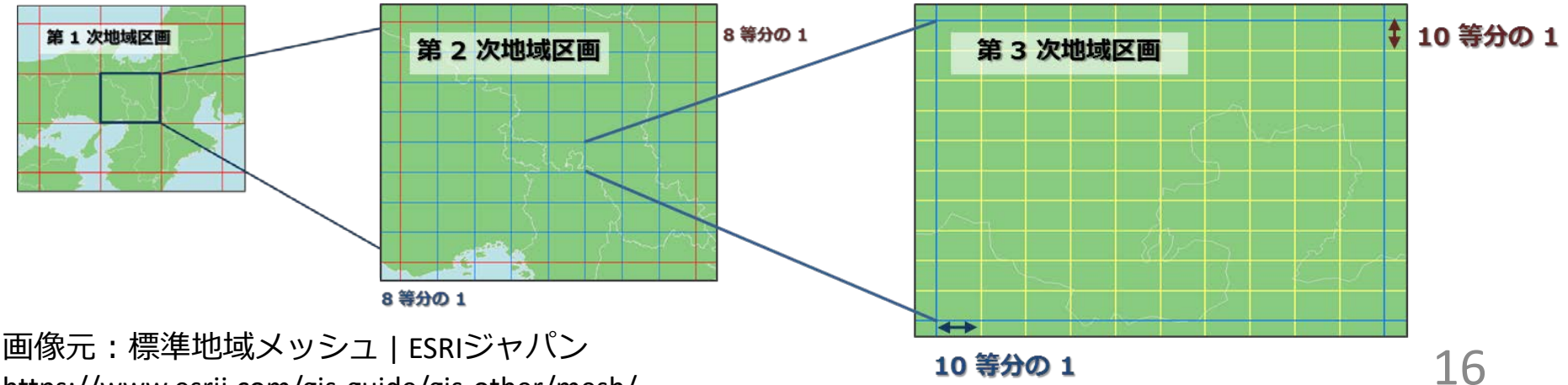
空間の分割（地域メッシュ統計）



地域メッシュ統計
に合わせて空間を分割

国勢調査等の結果との
位置合わせが簡単

３次メッシュは約1km四方

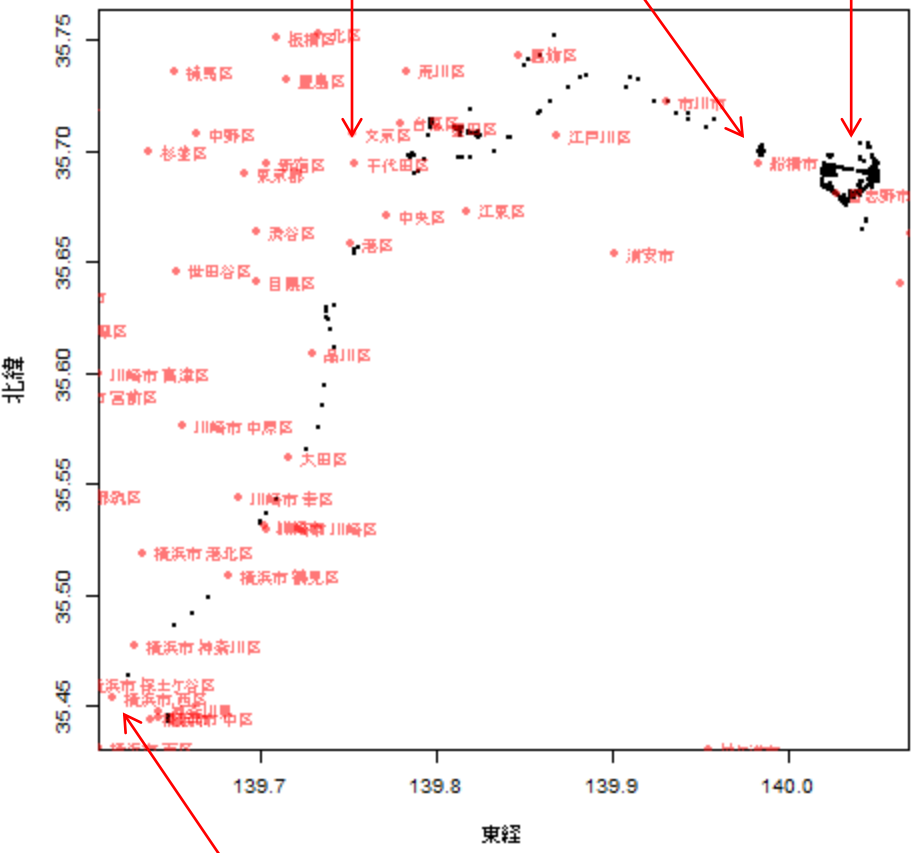


画像元：標準地域メッシュ | ESRIジャパン
https://www.esrij.com/gis-guide/gis-other/mesh/

ツイート拠点の判定

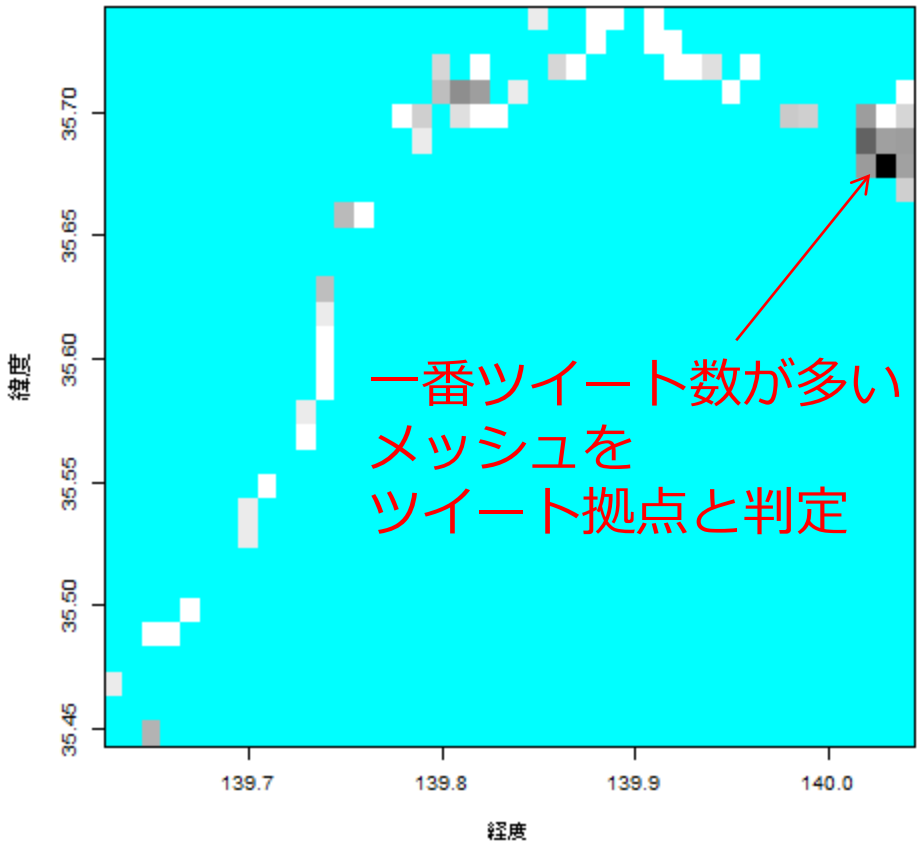
あるユーザーの位置情報付きツイート

千代田区　船橋市　習志野市



約1km四方

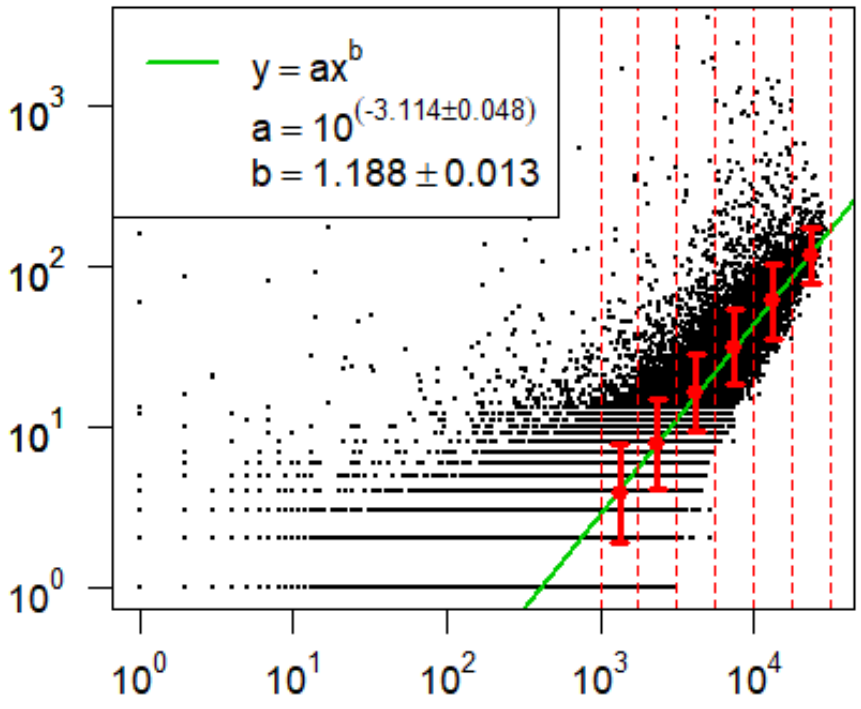
メッシュ（３次メッシュ）に切って
メッシュ内のツイート数をカウント



人口vsツイート拠点ユーザー数

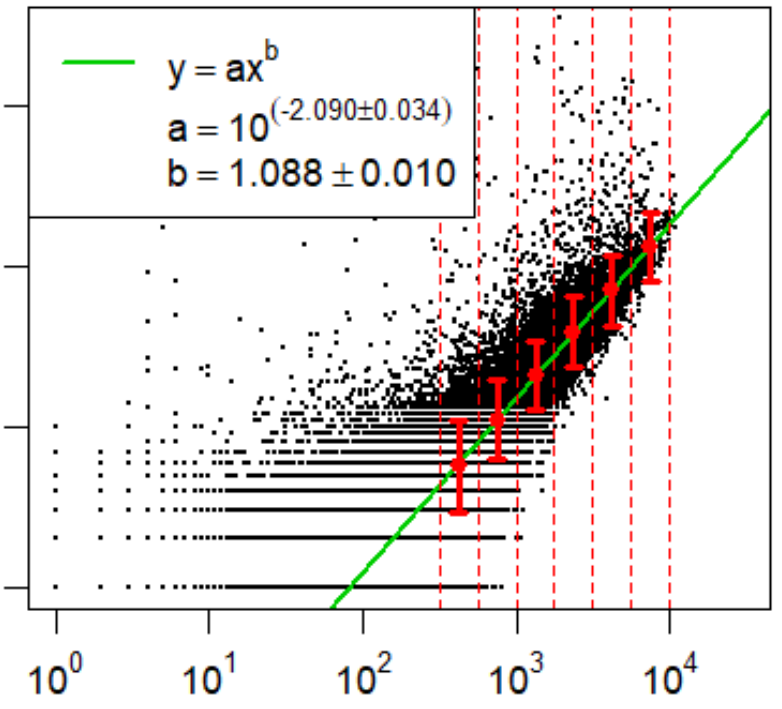
メッシュ内をツイート拠点としてるユーザー数

相関係数：0.46
対数データの相関係数：0.79



メッシュ内の人口

相関係数：0.49
対数データの相関係数：0.83



メッシュ内の20代30代人口

人口（住宅人口、夜間人口）と、ツイート拠点ユーザー数の比較結果

- 人口の大きなところは傾き1の直線
- エラーバーにx依存性はない
- 20代30代のツイッター利用者の多い人口でよりよく合う

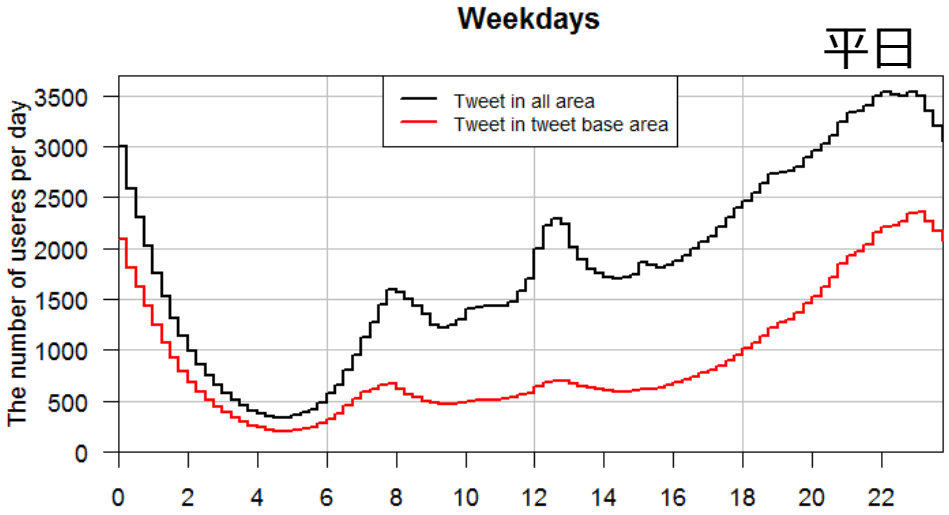
目次

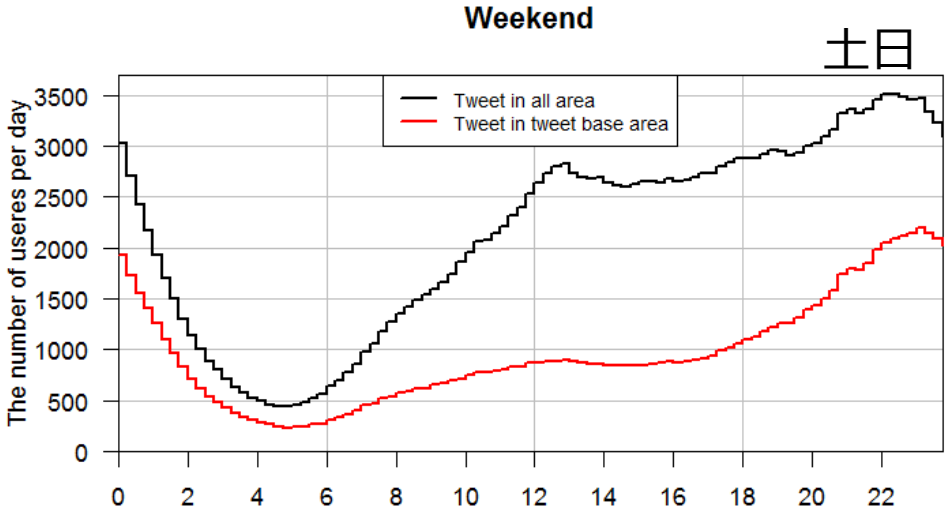
- イントロダクション
- データ
  - ツイッターデータ
  - 国民生活時間調査データ
- データ分析
  - 人口分布の再現
  - 起床在宅率の再現
- まとめと今後の課題

時間の分割

- 1時間を4等分（15分のビン）する
  - 比較対称の国民生活調査の結果に合わせた
- 15分間に一度でも発言したアクティブユーザー数をカウント
- このアクティブユーザー数を地域や曜日等の条件を付けて集計する

時間帯別
１日当たりのアクティブユーザー数

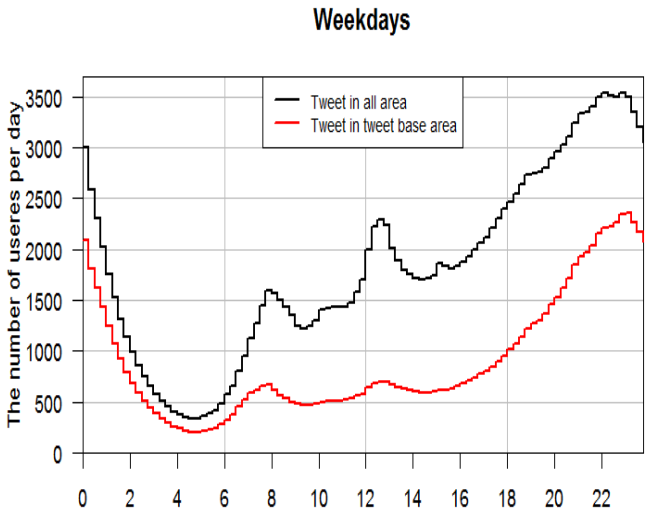


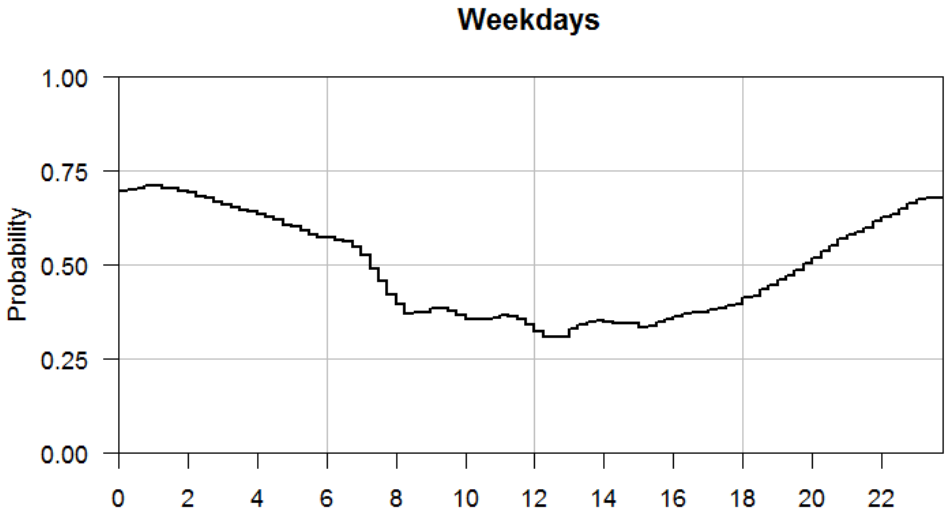


黒線：どこでツイートしたか
条件なしのユーザー数

赤線：ツイート拠点でツイートした
ユーザー数

ツイート拠点でのツイート率

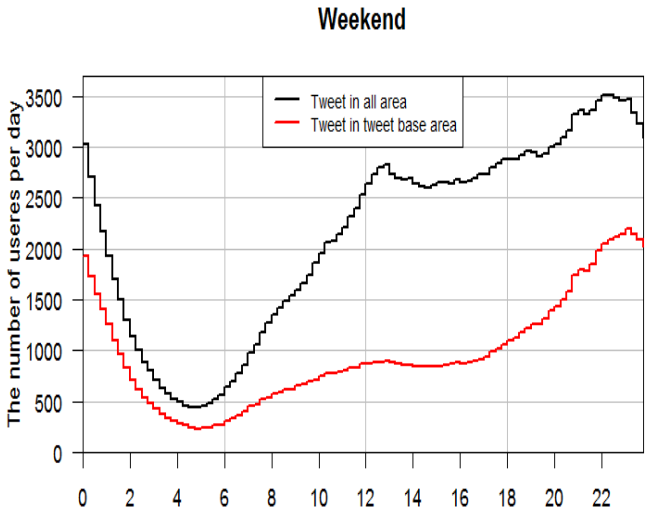


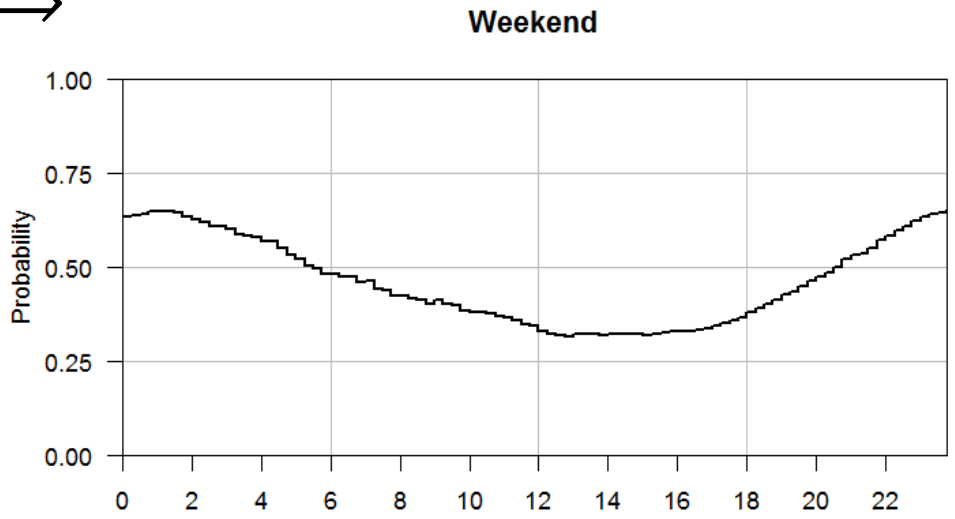


赤線

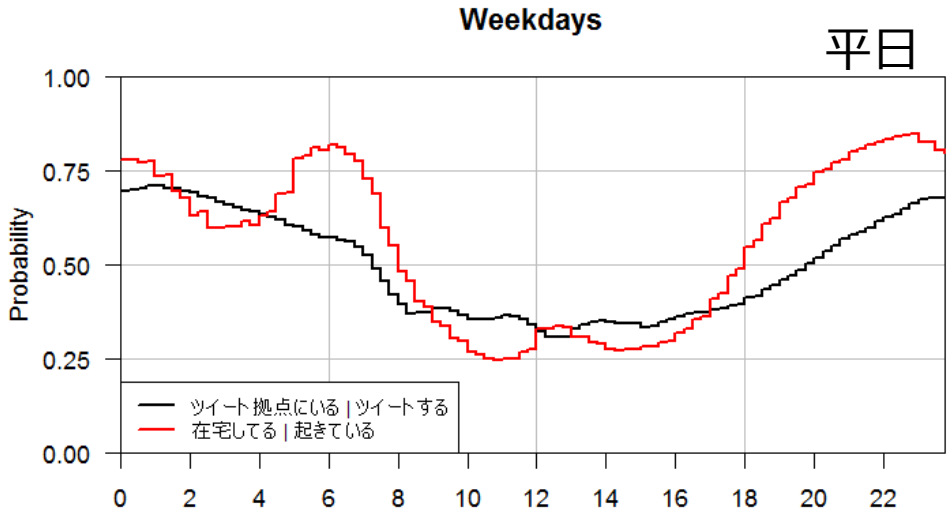
黒線

⇒



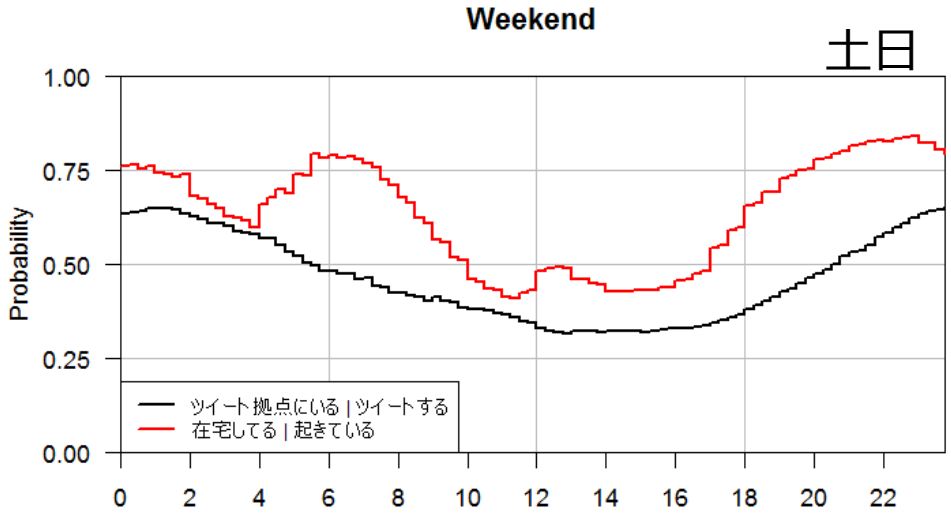


ツイート率と生活時間調査の比較



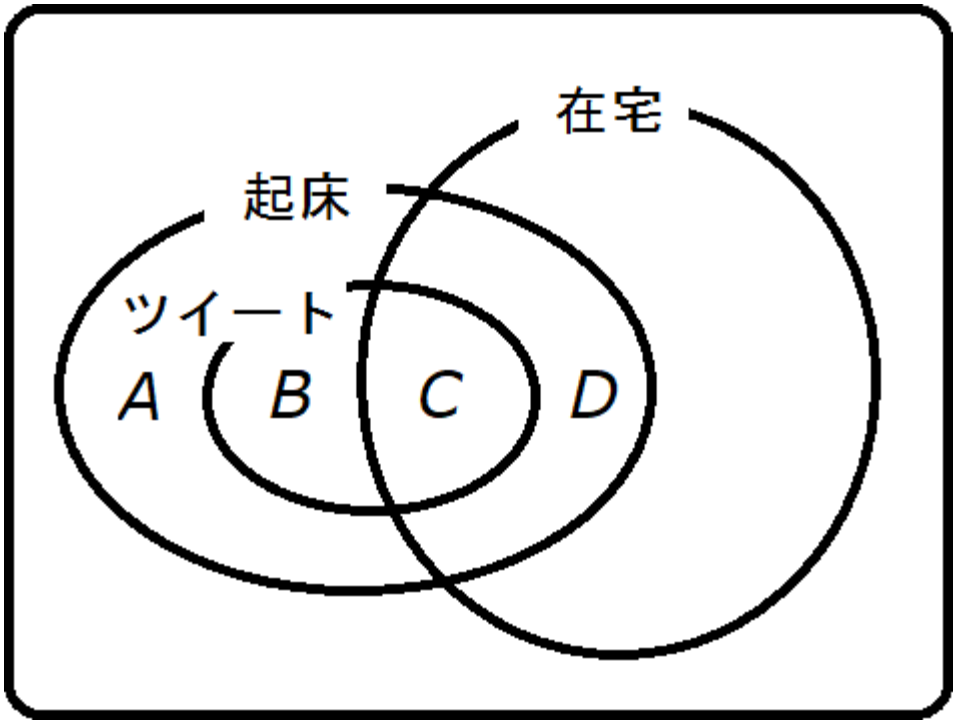
相関係数：0.864

黒色：*P*(ツイート拠点｜ツイート)
赤色：*P*(在宅｜起床)



相関係数：0.781

前頁の結果の解釈



ツイート拠点を自宅だと仮定

黒色：P(在宅｜ツイート)

$$= \frac{C}{B + C}$$

赤色：P(在宅｜起床)

$$= \frac{C + D}{A + B + C + D}$$

赤色＞黒色 の時（平日の昼や土日）は

非在宅での
ツイート/非ツイート

$$\frac{B}{A} > \frac{C}{D}$$

在宅での
ツイート/非ツイート

目次

- イントロダクション
- データ
  - ツイッターデータ
  - 国民生活時間調査データ
- データ分析
  - 人口分布の再現
  - 起床在宅率の再現
- まとめと今後の課題

まとめと今後の課題

- ツイッターデータの位置情報を用いて、国勢調査の人口分布と国民生活時間調査の起床在宅率が再現可能な事を示した
  - 頻繁な調査の可能性
  - 発展途上国での可能性
- ツイッターの付加情報を利用する
  - 人々の意識の違いによる分析
  - 防災意識とハザードマップとの関係