統計的機械学習による都市インテリジェンス研究

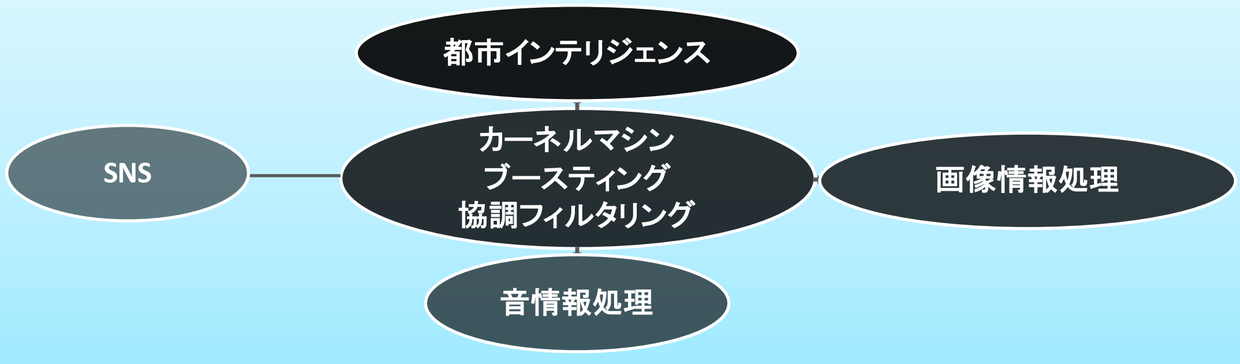
松井　知子　　モデリング研究系　教授

【概要】

本研究室では統計的学習機械を用いて、音声/音楽/画像/SNSなどを処理する方法について研究しています。具体的にはカーネルマシン、ブースティング、協調フィルタリングの手法を用いて、

1. 音声・話者認識
2. 音楽情報処理
3. 画像識別
4. SNS解析
5. WEBユーザビリティ評価
6. 都市インテリジェンス　など

の研究課題に取り組んでいます。



本研究室では統計的機械学習とその応用研究に興味のある学生さんを募集しています！

【統計的機械学習】

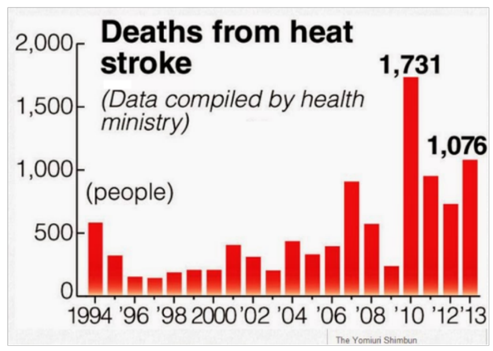
- 統計科学を用いて、
  - データから、内在する数学的な構造を発見する。
  - その数学的な構造に基づいて、予測や判別などの情報処理を行う。
- 帰納的アプローチ

  v.s.
- 自然科学でよく見られる演繹的アプローチ
  - 仮説をたて、推論し、実験的または理論的に検証する。

- カーネルマシン
  - 自動的な特徴(/モデル)選択機構を含む。
  - 非線形の扱いに優れている。
  - サポートベクターマシン(SVM)、罰金付ロジスティック回帰マシン
- いろいろな確率モデルによる方法
  - 混合ガウス分布モデル
  - 隠れマルコフモデル
- ガウス過程状態空間モデル　など

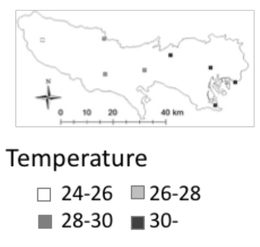
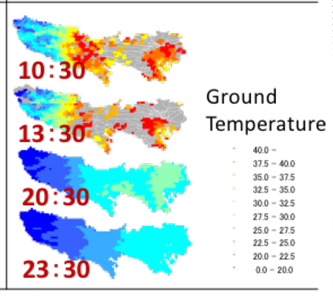
【都市リスク管理のためのツイートデータと異常気象事象の時空間解析】

Introduction



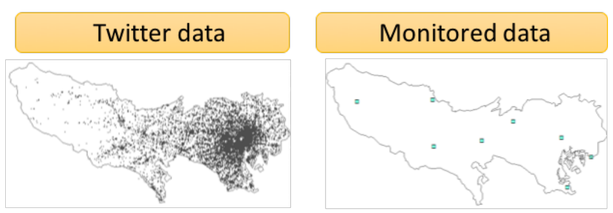
To cope with heatwave risks, **it is important to know the risk in a local (e.g., district-level) and real-time manner**.

Available temperature-related data

	Official monitoring data	Remote sensing data
Spatial resolution	Poor (8 monitoring stations)	High (1km grids)
Temporal resolution	Medium (every 1 hour)	Poor (10:30, 13:30, 20:30, 23:30)
Image	 Temperature □ 24-26 ▒ 26-28 ▓ 28-30 ■ 30-	

Twitter data

- Geo-tagged tweets in Tokyo in August 2012
  - They consist of 1% of random samples
  - Sample size (after a cleaning): 130,331



Extraction of heat-tweets

Heat-tweets : Tweets including any of synonyms of "heat"

あつい	Hot	蒸し	Muggy
熱い	Hot	水分補給	Rehydration
暑	Heat	体調管理	Health management
猛暑	Heat wave	猛烈に	Furious
炎天下	Blazing sun	だるい	Dull
真夏日	Hot day	死ぬ	Dying
残暑	Lingering summer heat	異常	Abnormal
熱中症	Heat illness	不快感	Discomfort
バテ	Faint	不快	Discomfort
寝苦しい	Cannot sleep well	イヤ	Unpleasant
夏本番	Midsummer	嫌	Unpleasant
日差し	Sunlight	クソ	Expletive
照り	Reflected heat	Orz	Expletive
湿度	Humid	きつい	Hard
湿気	Moisture	辛い	Hard
汗	Sweat	大変	Hard
ジメジメ	Damp	しんどい	Tired
ムシムシ	Humid	厳しい	Severe
ベタベタ	Sticky	苦手	Weak

Objective

As a first step of using tweets for a real-time and local heatwave management, we examine whether tweets are useful for local real-time temperature estimation.

- Research questions
  - **Do tweets explain temperature?**
    → A correlation analysis between heat-tweets and temperatures is conducted.
  - **How can we utilize tweets as a participatory sensing data that complement unobserved temperatures?**
    → The spatial best linear unbiased estimator (S-BLUE) is applied.

Model

- **Temperature**: $y(\mathbf{x}_i)$
  - Temperatures are describes by a latent spatiotemporal process, $f(\mathbf{x}_i)$, and a noise process, $v(\mathbf{x}_i)$.

$$y(\mathbf{x}_i) = f(\mathbf{x}_i) + v(\mathbf{x}_i)$$
$$f(\mathbf{x}_i) \sim N(0, c(\mathbf{x}_i, \mathbf{x}_j))$$
$$v(\mathbf{x}_i) \sim N(0, \sigma^2)$$

$\mathbf{x}_i$: Spatiotemporal coordinates of monitored temperatures
$c(\mathbf{x}_i,\mathbf{x}_j)$: Covariance between sites $\mathbf{x}_i$ and $\mathbf{x}_j$
σ^2: Variance parameter

- **Heat-tweets**: $y_{ht}(\mathbf{x}_I)$
  - It assumes that the probability of being a heat-tweet increases if $f(\mathbf{x}_I)$ exceeds a threshold temperature T.

$$P(y_{ht}(\mathbf{x}_I) = 1 \mid f(\mathbf{x}_I)) =$$

$$\sum_{p=0}^{1} P(y_{ht}(\mathbf{x}_I) = 1 \mid y_{ht\_0}(\mathbf{x}_I) = p) P(y_{ht\_0}(\mathbf{x}_I) = p \mid f(\mathbf{x}_I))$$

$$y_{ht\_0}(\mathbf{x}_I) = \begin{cases} 1 & if \quad f(\mathbf{x}_I) \geq T \\ 0 & otherwise \end{cases}$$

$\mathbf{x}_I$: Spatiotemporal coordinates of I-th tweet
$\mathbf{y}_{ht}(\mathbf{x}_I)$: 1 if I-th tweet is a heat-tweet, and 0 otherwise

Spatiotemporal temperature estimation equation

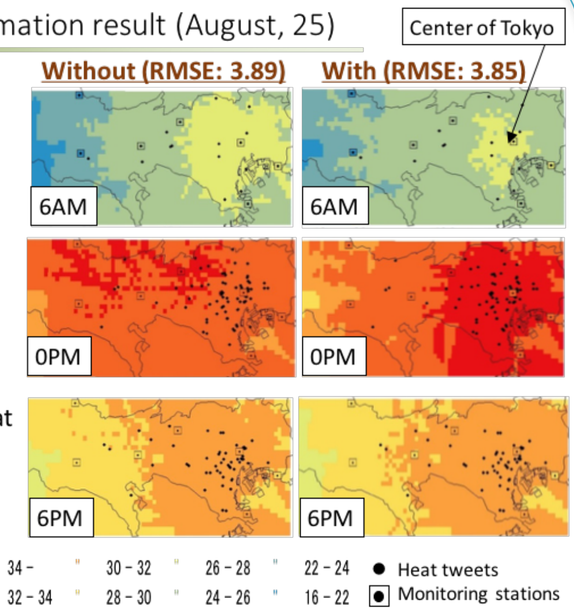
- **Spatial best linear unbiased estimator (S-BLUE)**
  - It estimates the temperature at $\mathbf{x}_*$, by minimizing the error variance.

$$\hat{f}_* = \hat{\alpha} + \hat{\mathbf{B}}\mathbf{Y} = \arg\min_{a,\mathbf{B}} E\left[(f_* - (\alpha + \mathbf{B}\mathbf{Y}))^2\right] = E\left[f_*\mathbf{Y}'\right]E^{-1}\left[\mathbf{Y}\mathbf{Y}'\right]\mathbf{Y}$$

f_*: The value of the temperature process at site $\mathbf{x}_*$
α: Constant
$\mathbf{Y}$: A vector of monitored temperatures, $y(\mathbf{x}_i)$ and heat-tweets, $y_{ht}(\mathbf{x}_i)$
$\mathbf{B}$: A matrix of unknown parameters

- $c(\mathbf{x}_i, \mathbf{x}_j)$ in $E[f_*\mathbf{Y}']$ and $E[\mathbf{Y}\mathbf{Y}']$ is usually given by a kernel function. However, the **kernel approach** is computationally expensive.
- Hence, we developed an EM-algorithm-based approach, which lighten the problem by defining $c(\mathbf{x}_i, \mathbf{x}_j)$ using radial **basis functions**.

Temperature estimation result (August, 25)



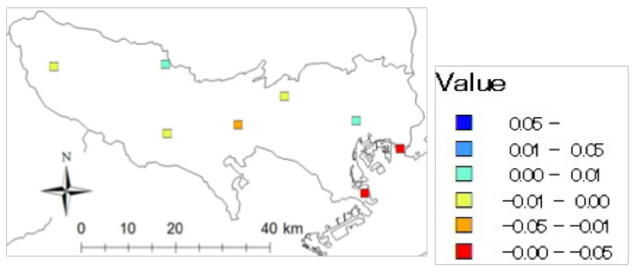
RMSE is improved by using tweets.

"With" approach successfully captures the heat island effect in the central area at the noon.

The strange result of "Without" approach at 0PM would be due to the small number of the monitoring stations (8).

Gap of RMSEs (With minus Without)

- Red : With twitter approach is better
- Blue: Without twitter approach is better



- The accuracy is improved in **the bayside area**
  - It might be due to the fact that heat-tweets capture the heat island effects in this area around noon.

Remaining issues

- Improvement of the heat-tweet extraction approach
  - Some words might explain heat stronger than others
  - Some tweets might not comment about the current situation, but about future prediction.
- Difference between feeling temperature and actual temperature
  - To analyze heatwave risks, it is needed to analyze feeling temperature, which depends on humidity, sunlight, mood/sentiment, etc.
  - Use of vital sensors would be needed.
- Fusion with other data
  - Remo sensing data, data of vital sensors etc.
- The computational complexity of S-BLUE
  - S-BLUE still requires the inversion of a covariance matrix, which becomes large if large samples are utilized.

共同研究者：
山形与志樹、村上大輔(国立環境研究所)
Gareth W. Peters(UCL)