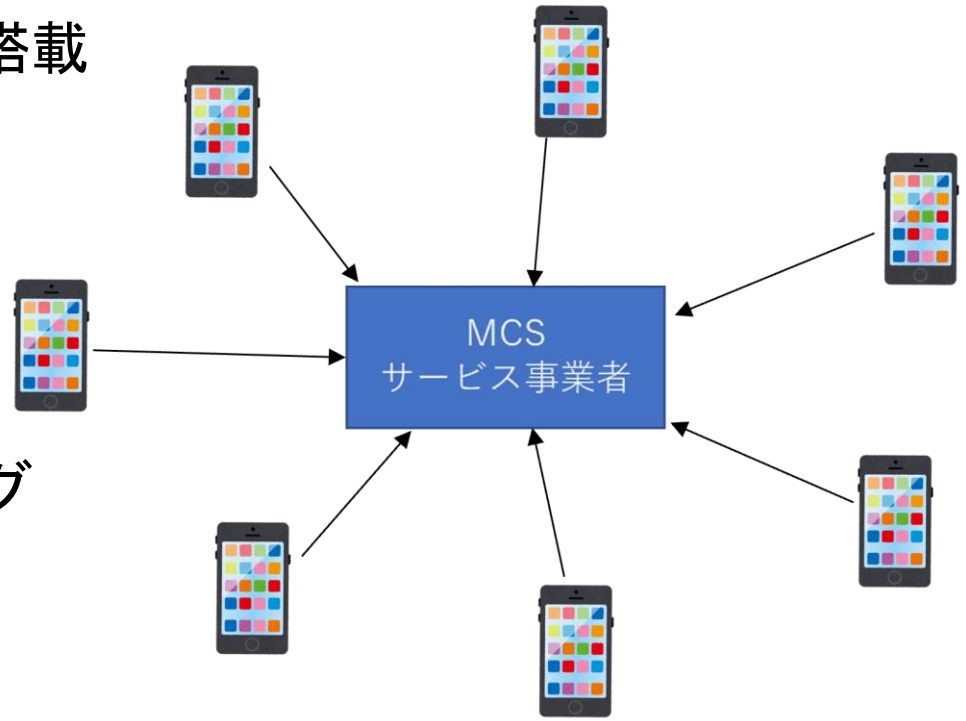
複数地域の Crowdsensing における ワーカの最適サンプリング

Optimum Worker Sampling in Crowdsensing with Multiple Areas

松浦千紘　上山憲昭
立命館大学 情報理工学部

研究の背景

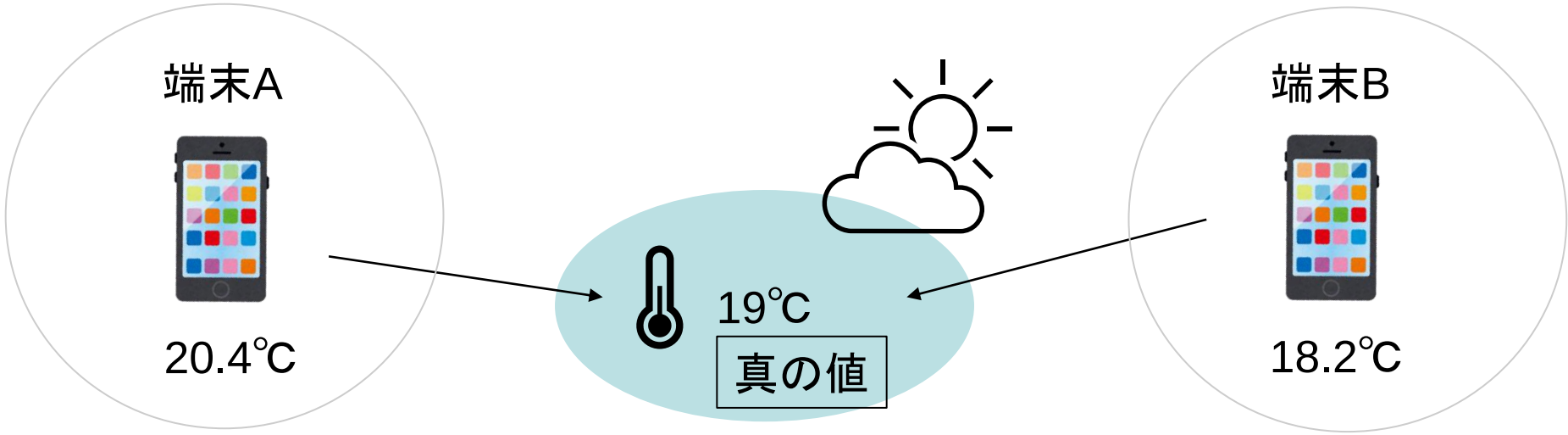
- モバイルクラウドセンシング（MCS：mobile crowdsensing）
  - モバイル端末をIoTデバイスとして活用
  - 今日のモバイル端末は様々なセンシング機能を搭載



- MCSの利点
  - インフラの新規構築が不要なため低コスト
  - 従来のIoTデバイスに劣らない高機能なセンシング
  - 普及率が高く膨大なデータを収集可能

- 活用例
  - サービス事業者がワーカからデータを収集，真の値を推定
  - 気象予測などに役立てる

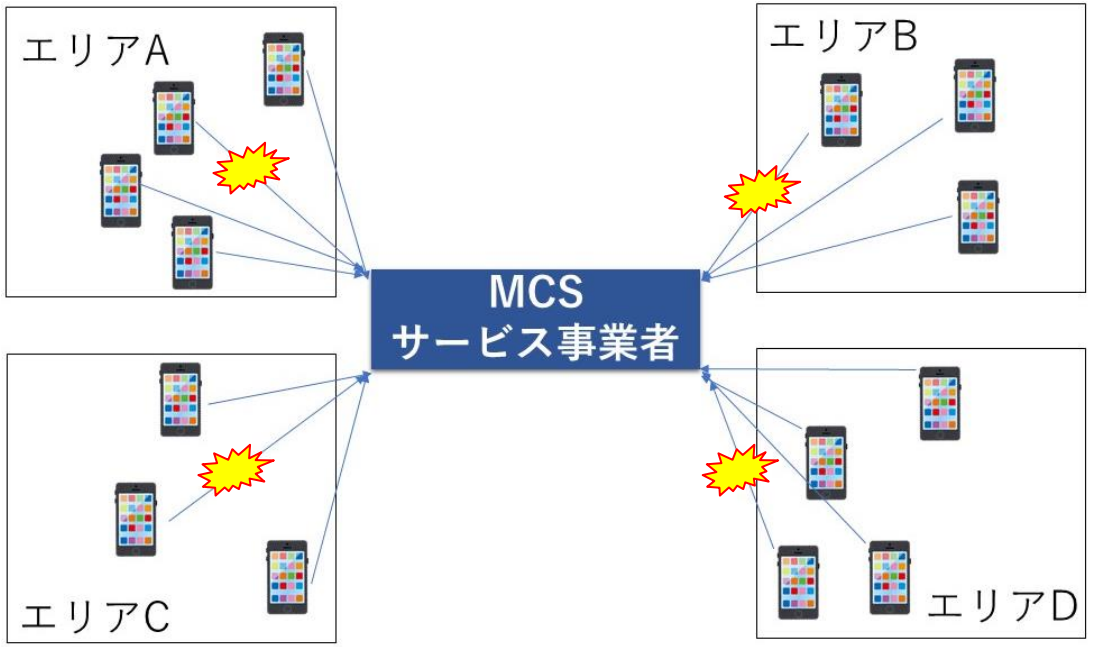
MCSの問題点



- 計測時の誤差の発生
  - センサーの不具合やヒューマンエラーによった誤ったデータの発生
  - 悪意のあるワーカによる誤ったデータの送信
    → 推定値が歪むデータポイズニング攻撃の発生

関連研究

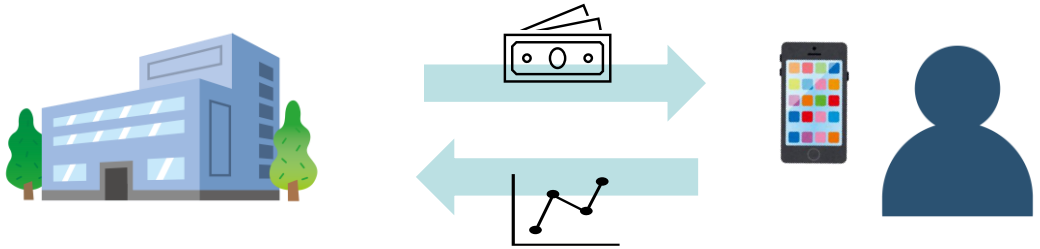
- ワーカ推定誤差を最小化するよう，各ワーカの測定値を重みづけした重みづけ平均で推定する CRH（Conflict Resolution on Heterogeneous data）法の提案 [1].
- 複数エリアごとに複数ワーカから測定値を推定する MCS において，攻撃ワーカが推定誤差を最大化するように各エリアの配置攻撃者数を最適化する方式の提案 [2].



[1] Q. Li, et al., Conflicts to Harmony: A Framework for Resolving Conflicts in Heterogeneous Data by Truth Discovery, IEEE Trans. Know. Data Eng., 28 (8), Aug. 2016
[2] R. Fujimoto and N. Kamiyama, Poisoning Attacks in Crowdsensing Over Multiple Areas, IEEE GLOBECOM 2022

研究の目的

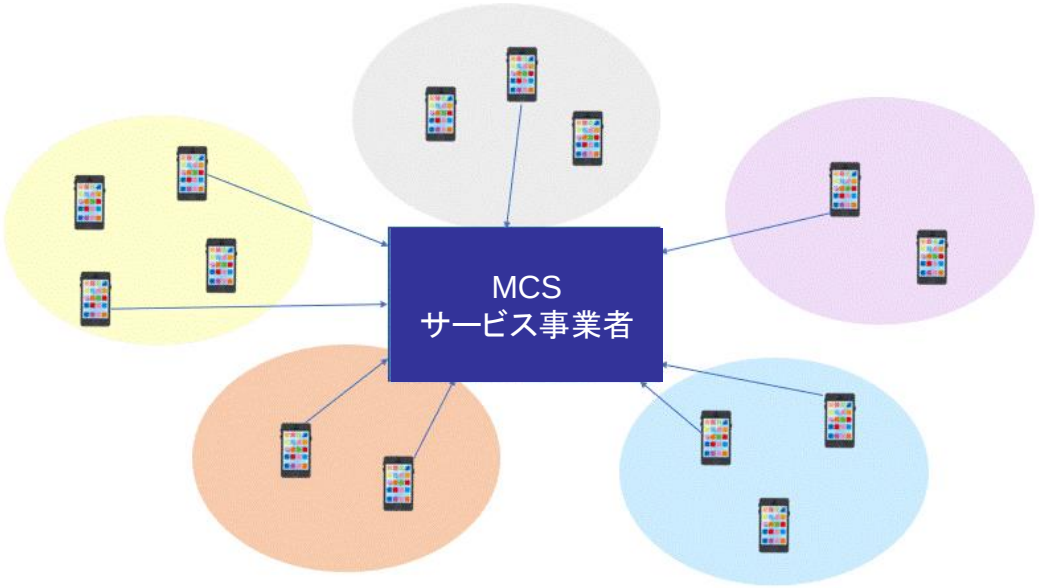
- 着目する課題
  - データ収集時，ワーカにはインセンティブの提供が必要　図　入れる
    → 一定の確率でサンプルしたワーカからのみデータを収集



複数エリアのMCSにおいて，全エリアの誤差総和の最小化を目的とした
各エリアの最適サンプル数設定法の提案

複数エリアMCS

- データ収集領域は複数の K 個のエリアから構成
- 各エリアにおいてセンシングデータをモバイル端末のユーザから収集し，エリアごとに測定値を推定
  - 市街を 100 メートルの正方領域に分割し，各エリアの二酸化炭素濃度を推定
- サービス事業者が収集できる総サンプルワーカ数をNとしたときの，各エリアのサンプルワーカ数を決定する問題を考える



CRHアルゴリズム(Conflict Resolution on Heterogeneous data)

- 目的
  - 複数の測定値から真の値を推測する

- 概要
  - 真の値と測定値との差異が小さいワーカの信頼性は高く，大きいワーカの信頼性は低くなるように各ワーカの信頼性を設定
  - 信頼性を重みとした測定値の加重平均を推定値として用いる
  - 推定値と各ワーカの重みを交互に更新し，収束した値を推定値として扱う

- アルゴリズム
  1. 各ユーザ k の信頼性（重み）w_k を 1 に初期化
  2. 式(1)で，各ユーザ k の測定値 v_k と w_k から推定値を計算
  3. 式(2)でユーザ毎の信頼性 w_k を更新
  4. 推定値及び信頼性が収束するまで step 2, 3 を反復

$$w_k = -\log \frac{(v_k - \tilde{v})^2}{\sum_{k \in N \cup A} (v_k - \tilde{v})^2} \quad (1)$$

$$\tilde{v} = \frac{\sum_{k \in N \cup A} v_k w_k}{\sum_{k \in N \cup A} w_k} \quad (2)$$

（正常ワーカの集合を N，攻撃ワーカの集合を A とする）

提案方式で使用するパラメータ

- 提案方式で用いるパラメータを以下の表に示す

記号	定義
K	エリア数
N	総サンプルワーカ数
u_i	各エリア i のサンプルワーカ数
p_i	各エリア i の測定値の真値
μ_i	各エリア i のワーカ報告値の平均値
σ_i	各エリア i のワーカ報告値の標準偏差
v_i	エリア i の推定値
E	総推定誤差
$e_{i,ui}$	サンプルワーカ数 u_i の場合のエリア i の推定誤差
S_0	初期配置状態
S_1	サンプリング数最適化後の配置状態
η	総誤差の差分の判定に用いる閾値

提案方式(1/2)

- 目的
  - 総サンプルワーカ数の上限 N を制約条件として考慮し，本条件下で総誤差 E が最小となるよう各エリア i のサンプルワーカ数 u_i を最適設計

- 最適化問題
  - エリア i の推定値を v_i ，真の値を p_i とすると，各エリアの誤差は $|v_i - p_i|$ と表されるので
    目的関数を（1）式のように立式（K:エリア数）

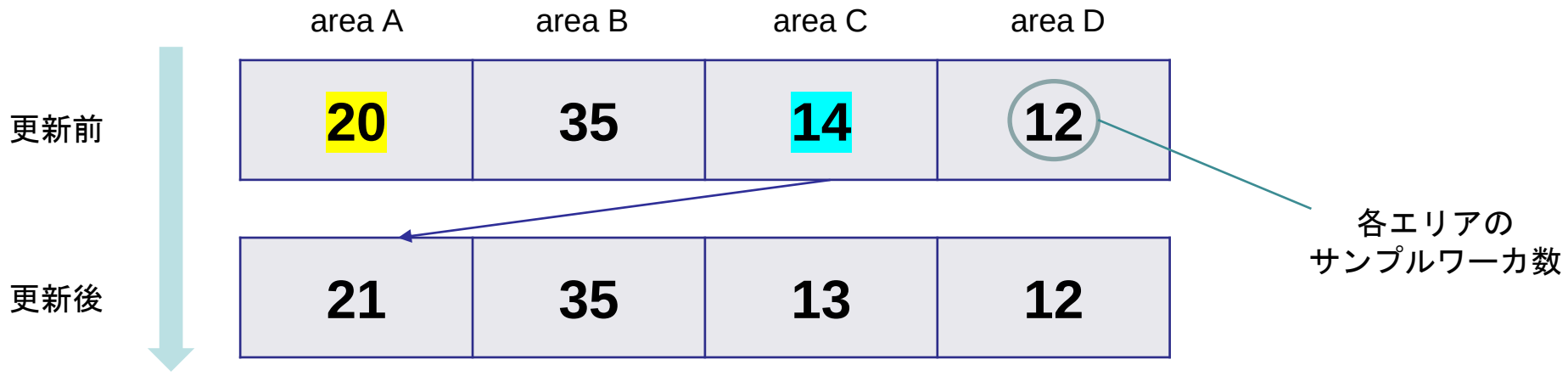
$$\min E(u_1, u_2, \cdots, u_K) = \sum_{i=1}^{K} (v_i - p_i)^2 \qquad (1)$$

  - 制約条件を（2）式に示す

$$\sum_{i=1}^{K} u_i = N \qquad (2)$$

提案方式(2/2)

- アルゴリズムの概要
  1. 各エリアのサンプルワーカ数の初期値を $u_i = N/K$ に初期化し，このときの総誤差 E_{ini} を算出
  2. ランダムにサンプルワーカ数を与えたときの平均推定誤差を算出し，各エリアのサンプルワーカ数に対する平均推定誤差 $e_{i,ui}$ の近似解を得る（DBに格納）
  3. 各エリア i のサンプル人数をインクリメント（デクリメント）し，推定誤差の減少量 e_{dec} (増加量 e_{inc}) を算出
  4. 減少量が最大であるエリアのサンプル人数をインクリメント，増加量が最小であるエリアのサンプル人数をデクリメント
  5. 総誤差の変化量 $|E_{post} - E_{pre}|$ が閾値 η を下回るまでこれを反復し，このときの総誤差 E_{conv} を算出

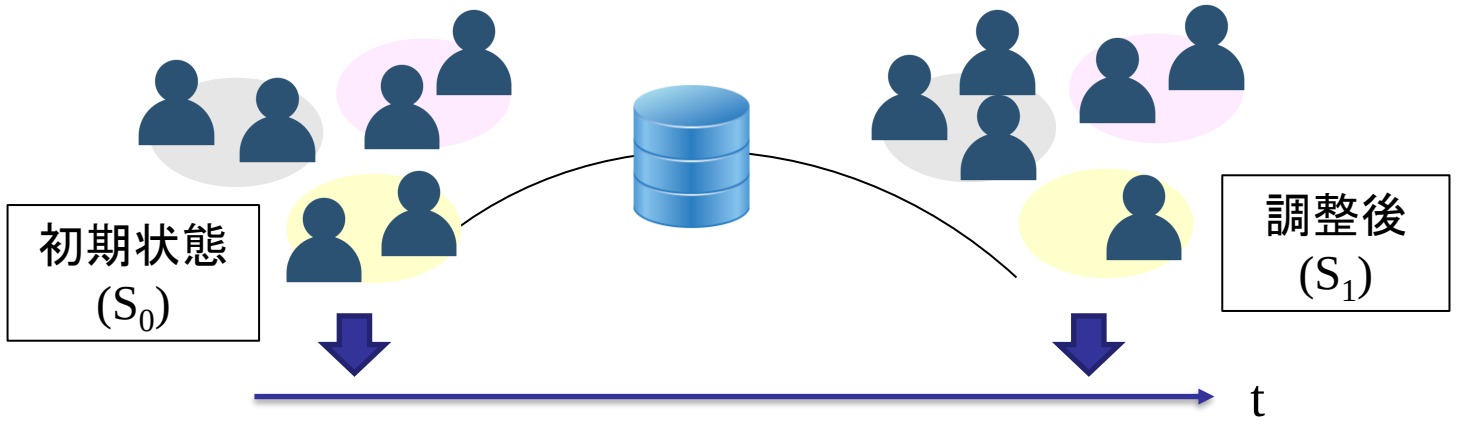


実験条件

- 数値条件

記号	値
K	10
N	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000
μ_i	50
σ_i	2, 7, 12, 17, 22, 27, 32, 37, 42, 47
η	10^{-5}

- シミュレーション条件
  - サービス事業者は時刻 t_n で推定値を計算
  - 正常ワーカのみ（調整前：t_0，調整後：t_1）の2つの状態を設定

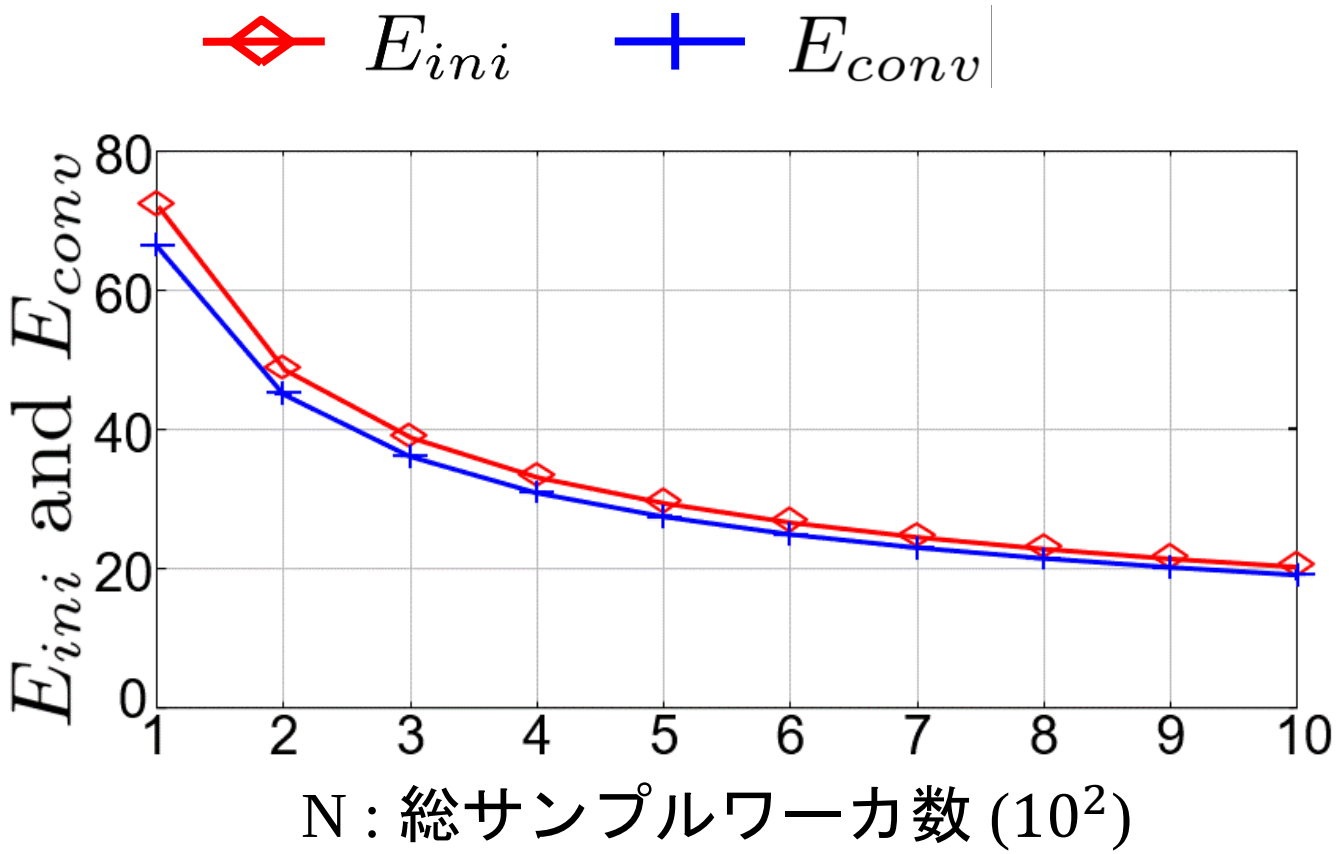


サービス事業者は観測時刻に
母集団からデータを提供する
ワーカをランダムに選出

サービス事業者が推定値計算を
行う時刻

評価結果

E_{ini} : 初期状態 S_0 の総推定誤差　　E_{conv} : 調整終了後 S_1 の総推定誤差



- 提案方式適用後の E_{conv} は E_{ini} を下回り，総推定誤差抑制効果を確認
- 同サンプルワーカ数における総誤差の減少量（$E_{ini} - E_{conv}$）はサンプルワーカ数の増加に伴い減少
- ワーカの測定値の標準偏差が大きいエリアにより多くのワーカが配置

まとめ

- 本研究では，総推定誤差の最小化を目的として各エリアの最適サンプル数設定法を提案
  - 総サンプルワーカ数が固定であるという条件のもと複数エリアからワーカのデータを取集
  - ワーカの測定値の分散が最適サンプル数や推定精度に与える影響を明らかにした

- 今後の方針
  - 防御法
    サービス事業者がワーカの区別を行うことができるモデルを想定
    提供するインセンティブを考慮したサンプリング方法の考案
  - 攻撃法
    上記の条件のもと，攻撃ワーカの効果的な配置方法を考案