2022年度　統計データ分析コンペティション

# 審査員奨励賞［大学生・一般の部］

# IT社会が生み出す連鎖する格差

山邊 璃久（信州大学大学院総合理工学研究科）

# IT 社会が生み出す連鎖する格差

山邊 璃久[*1]

*1: 信州大学大学院 総合理工学研究科 工学専攻 2 年

## １．研究のテーマと目的

現在の生活の中で、インターネットは必要不可欠な存在となっている。情報の収集、商品購入、オンライン学習など、様々な領域で使用されるためインターネットを利用する人と利用しない人とでは生活に大きく差が出てしまう。特に、地方と都心ではインターネット利用率に大きな差が生まれる、情報格差（デジタルディバイド）が発生している。情報格差が引き起こす問題は、情報を収集できないという単純な問題だけでなく、高齢者の孤立、緊急時の対応遅れ、所得の格差といった社会問題にも発展している。この問題が発生している原因として、マイノリティー（情報端末を利用しない人々）への対応不足が考えられる。多くの人が利用しているものを基準に技術や施策の進化が起こるため、マイノリティーの人々への対応は次第におろそかになってしまう。その対応不足が重なり、多くの不便が生じてしまい、社会との距離が広がっていってしまう。したがって、マイノリティーへの対応は今後の課題の１つとして非常に重要である。マイノリティーへの対応を検討するにあたり、明確化しなければならないのがインターネット利用者と非利用者の特徴である。両者の違いが明確化されれば、的確なアプローチで情報格差を無くすことができる。情報格差による研究(1)では、4 種類の格差があると示した。個人間の格差、組織間の格差、地域間の格差、国際的な格差である。この 4 種類の原因をそれぞれ特定し、それぞれの格差によってどんな問題が生じるかを示すことができれば、問題が発生する前に施策を打てる。本研究では４つの格差のうち、地域間の格差を分析し、今後発生しうる問題について検討していく。地域の格差が無くなれば、組織の格差がなくなる。組織の格差がなくなれば、個人の格差もなくなる。このように連鎖的に格差を解消できると考えている。したがって、地域の格差を分析することが重要である。本研究では、この分析にあたり Random Forest の特徴量重要度を活用した。特徴量重要度を基に、仮説を立て、特徴量を可視化することにより、地域の格差により今後引き起こされる課題を検討した。

## ２．研究の方法と手順

本研究では、SSDSE データセットに都道府県ごとのインターネットの利用率のデータ(2)（以下、IT 利用率とする）を加え、IT 利用率によって起こる変化を調査していく。以下に詳細な調査手順を示す。

### 2-1. データセット作成

使用するデータセットについては、主に 2 種類使用する。（１）SSDSE の社会生活データ（SSESD-D）（２）都道府県別インターネットの利用率(2)この２つのデータセットの組み合わせ方や加工方法は第３節に記述する。また、データセットの可視化を行う際にのみ、SSDSE の基本データ（SSESD-A）の 1 人あたりの県民所得を使用する。

### 2-2. Random Forest による特徴量重要度の算出

IT 利用率の違いによって、地域間にどのような変化があるのかを調査するため Random Forest の特徴量重要度を用いる。Random Forest とは、複数の決定木を利用した学習モデルである。弱学習器を複数用いることで、性能を向上させることができるバギングを使用している。また、決定木は過学習しやすい傾向にあるが、複数の弱い決定木を組み合わせるため、精度を向上させることができる。特徴量重要度は Random Forest が

予測・分類する際に、それぞれの特徴量をどの程度重視したかを示す指標であるため、予測・分類対象に密接に関係している特徴量を探すのに活用できる。これらを使用するため、目的変数を IT 利用率、特徴量を SSDSE の社会生活データとして入力して特徴量重要度を取得する。評価指標は正解率とする。評価指標を定義したのは、予測性能がある程度高いことを証明することで特徴量重要度の信頼度を上げるためである。また、評価指標を正解率にしたのは、誰にとってもわかりやすい指標だからである。

### 2-3. インターネットの利用率によって起こる変化の特定

特徴量重要度が高いもの上位 20 個を抽出し、それぞれを IT 利用率と比較することで IT 利用による変化を特定する。主に、matplotlib を活用した棒グラフでの可視化を活用する。また、共通してみられるものがあればグループ化して考察していく。

## ３．データセットの加工

本研究で使用したデータセットは、2 種類のデータセットを組み合わせて作成している。種類については、2-1 で示したものである。以下に、詳細な加工手順を示す。

### 3-1. データセットの作成

（１）のデータセットを Random Forest に入力できるようにしていく。データ形式としては、データフレーム構造を使用する。データフレーム構造とは、最初の行に列の名称があり、列の名称に書いてあるデータの種類が同じ列に入っているデータ形式である。このようにして、列１～列４のように重ねたものをデータフレーム構造（表１）という。このような構造に変形するため、SSDSE から csv ファイルで取得し、1 行目を飛ばしつつ、Encoding を shift-jis とした。（２）のデータセットは、総務省の報告書[(2)]からデータ情報を確認し、csv ファイルを作成した。このデータを（１）のデータセットに結合して１つのデータセットとした。

Random Forest に入力する際には、カテゴリ情報を含む列を削除した。詳細は、表２に記載した。

表１：データフレーム構造のイメージ

| 都道府県 | 最高気温 | 最低気温 | 湿度 |
|---|---|---|---|
| 北海道 | ○○℃ | ○○℃ | ○○% |
| 青森 | ○○℃ | ○○℃ | ○○% |
| 秋田 | ○○℃ | ○○℃ | ○○% |

表２：削除する列名

| 列名 | 詳細 |
|---|---|
| 男女の別 | 男女別で分けるためのラベル |
| 地域コード | 地域ごとに決められた特定の番号 |
| 都道府県 | 都道府県の名称 |
| 起床（平日の平均時刻） | 起床の時間 |
| 朝食開始（平日の平均時刻） | 朝食の時間 |
| 夕食開始（平日の平均時刻） | 夕食の時間 |
| 就寝（平日の平均時刻） | 就寝の時間 |
| 出勤（有業者、平日の平均時刻） | 出勤の時間 |
| 仕事からの帰宅（有業者、平日の平均時刻） | 帰宅の時間 |

### 3-2. IT 利用率の Bin 分割

都道府県ごとの IT 利用率の Bin 分割を行い、利用率ごとに高・中・低の３つに変換した。行った理由としては、IT 利用率ごとの地域の特徴を明確化したかったため回帰問題で IT 利用率を予測させるよりも分類問題で IT 利用率ごとのグループを分類した方が特徴量重要度を参考にできると考えたからである。Bin 分割をどのように行ったかは、以下に示す。

表３：Bin 分割の詳細

| 分割基準 | 分割後のラベル |
|---|---|
| 0 ＜ IT 利用率 ≦ 85 | 0 |
| 85 ＜ IT 利用率 ≦ 90 | 1 |
| 90 ＜ IT 利用率 ≦ 100 | 2 |

### 3-3. データセットの標準化

データセットに含まれる値の単位などが異なるため、標準化することで同じ尺度で異なる特徴量を比較することができる。標準化については、以下のように行う。n は、全データ数。x は、各データの１つの値とする。簡単に流れを説明していく。まず、データの平均値（$\bar{x}$）を求める。その後、平均値を活用して、標準偏差（$\sigma$）を求める。標準偏差と平均値から標準化（$x_i'$）を行う。Python のライブラリには、この計算を自動で行う Standard Scaler 関数があるため、今回はそれを使用する。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$x_i' = \frac{x_i - \bar{x}}{\sigma}$$

### 3-4. 学習・評価用のデータ分割について

Random Forest で学習・評価するにあたりデータを分割する必要がある。学習に使用したデータを評価で使用すると正しく予測できたのか、過学習しているだけかを判断できないため、正しくデータ分割することが重要である。本研究では、全データのうち 60%を学習用とし 40%を評価用とした。また、学習用と評価用に分ける際に、目的変数のラベル（IT 利用率の高・中・低のラベル）の割合を同じにした。これにより、学習と評価において同様の条件で判断することできる。

## ４．データ分析の結果

研究の手順に従い、データ分析した結果を以下に示していく。

### 4-1. Random Forest による特徴量重要度の算出

Random Forest により、学習を行った。学習の際に使用した Random Forest のパラメータは、デフォルトのままである。パラメータ最適化をすると、来年以降のデータに当てはめた際に性能が下がる可能性がある。よって、来年以降に同じアプローチで検証することが難しくなってしまい、比較検討できないので最適化は行わ

なかった。学習後の分類結果では、評価データにおいて正解率 73.68%を得た。機械学習を使用しないでランダムに分類すると 50%のため、73.68%の分類精度は高いと考えている。したがって、IT 利用率をきちんと分類できるため、特徴量重要度も信頼性がある。表 4 は特徴量重要度の上位 20 個である。この重要度では、趣味が多く入っている。つまり、IT 利用率の変化により、その趣味を行う人に傾向が生じているということである。使用したデータは、生活の中で何をしている時が多いかといった生活データである。言い換えると、休日にどのようなことをしているかが分かるデータといえる。休日の過ごし方は、大きくわけると「積極的休養」と「消極的休養」に分かれる。IT 利用率によって、休養タイプに偏りが生じるかを次の節で可視化してみていく。

表 4　特徴量重要度の上位 20 個

| 順位 | 列名 |
|---|---|
| 1 | 楽器の演奏 |
| 2 | スポーツ（総数） |
| 3 | パチンコ |
| 4 | CD・スマートフォンなどによる音楽鑑賞 |
| 5 | 詩・和歌・俳句・小説などの創作 |
| 6 | ウォーキング・軽い体操 |
| 7 | 映画館以外での映画鑑賞 |
| 8 | キャンプ |
| 9 | 休養・くつろぎ |
| １０ | 障がい者を対象とした活動 |
| １１ | まちづくりのための活動 |
| １２ | 英語 |
| １３ | 卓球 |
| １４ | 外国語 |
| １５ | 商業実務・ビジネス関係 |
| １６ | 通勤・通学 |
| １７ | 交際・付き合い |
| １８ | テレビゲーム・パソコンゲーム |
| １９ | 登山・ハイキング |
| ２０ | カラオケ |

### 4-2. インターネットの利用率によって起こる変化の特定

IT 利用率が高い県と低い県でどのような変化があるのかを可視化して検討していく。ここでは、IT 利用率の高い県と低い県のみ表示する。両極端なグループを見る方が検討しやすいためである。各グラフでは、縦軸をデータ数（標準化された値であることに注意）、横軸を都道府県の名称とする。

図 1 では、楽器の演奏をする人と IT 利用率の比較を行っている。楽器の演奏をする人と IT 利用率には正の相関があることが分かる。図 2 では、スポーツをする人数と IT 利用率の比較を行っている。スポーツをする人も IT 利用率と正の相関があることが分かる。同様に、他の趣味においても IT 利用率との正の相関が見

受けられた。以上のことから、IT 利用率の高い県では、休日には自分から積極的に動き楽しいと思えることを行う「積極的休養」を行う傾向にあることがわかる。一方、図３の１日で休養・くつろぎに時間を使う割合と IT 利用率の比較を見ると、休養・くつろぎをする人と IT 利用率は負の相関になっている。つまり、IT 利用率の低い県では、休みの日や空いている時間には、あまり動かずおとなしくしている「消極的休養」をすることが多いと考えられる。これは、趣味をする場所や機会が都市には多く、地方には少ないことも関係している。都市の方が、習い事教室の数や専門店が多いため趣味を始めやすい。一方、地方では教室も専門店も少ないため、趣味を始めるのも難しいことがある。スマートフォンの登場により、情報へのアクセス権が平等になることで地方と都市の差がなくなることが期待されたが、店の立地等の理由で休日の休養方法の差を無くすことはできなかった。

図４では、1 人あたりの県民所得と IT 利用率の比較を行っている。県民所得と IT 利用率には正の相関があることが分かる。IT 利用率が高い県は積極的休養を行い、IT 利用率が低い県は消極的休養を行っている。この違いが、県民所得に影響を及ぼしている可能性も考えられる。理由としては、県民所得を生み出しているのは、県民の労働だからである。そこで、5 節では休養方法の違いにより生じる変化について検討していく。

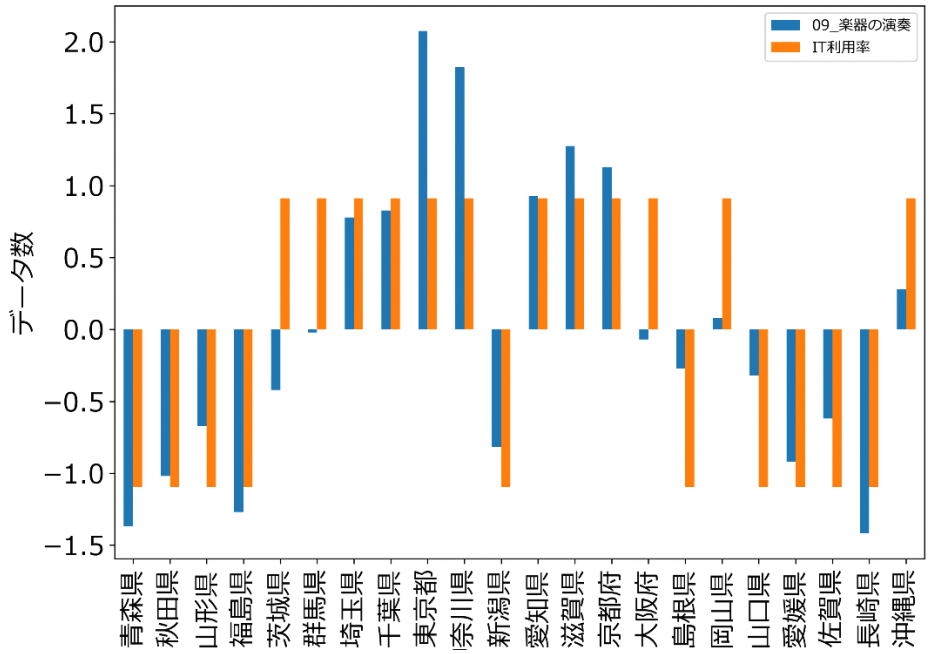


図１：楽器の演奏をする人数と IT 利用率の比較図

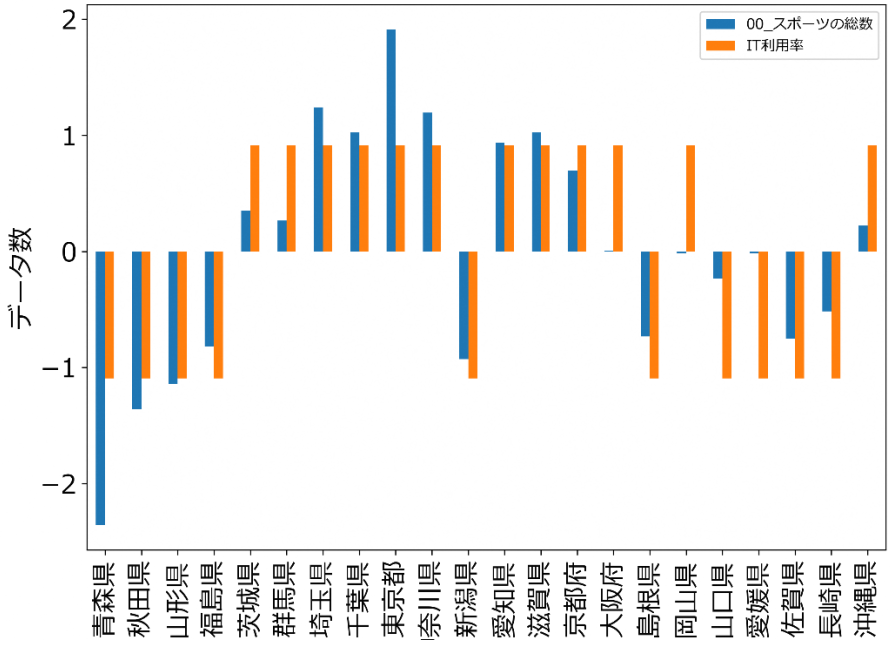


図２：スポーツをする人数と IT 利用率の比較図

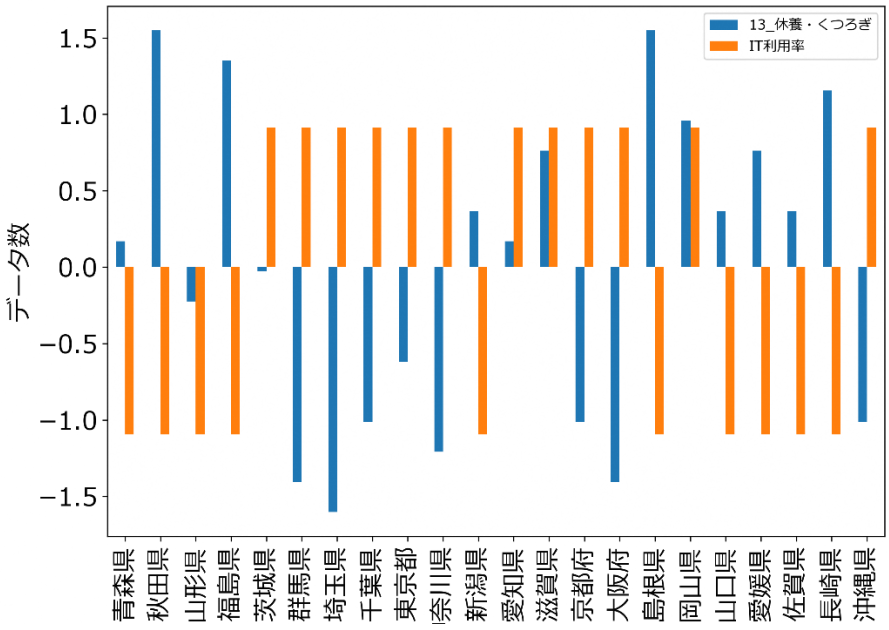


図３：1 日で休養・くつろぎに時間を使う割合と IT 利用率の比較図

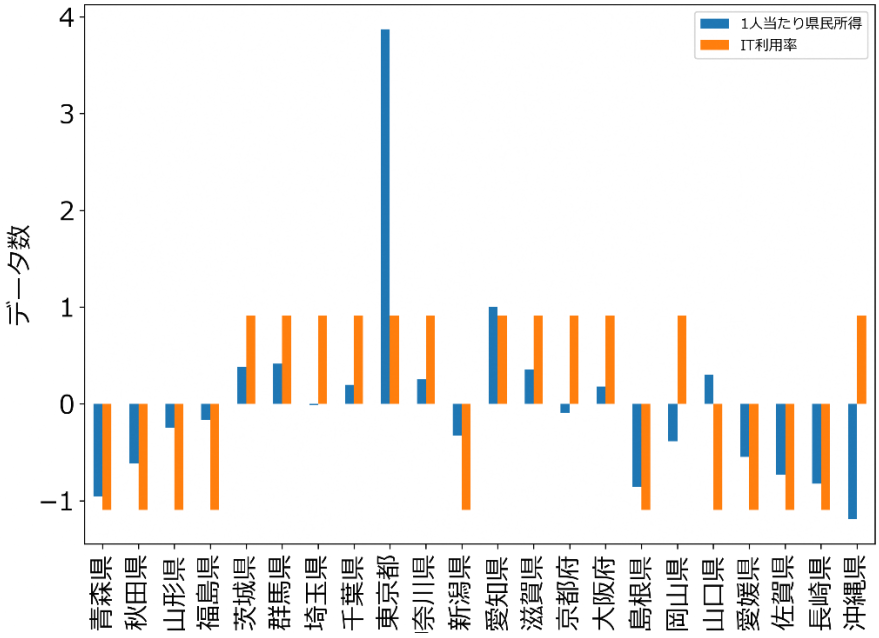

図４：1 人あたりの県民所得と IT 利用率の比較図

## ５．結果の解釈

IT 利用率と特徴量の比較から、休養方法に違いがあることがわかった。この違いにより、どんな変化が引き起こされるのかを検討していく。

### 5-1. 休養方法の違いによるパフォーマンスの変化

IT 利用率の変化と特徴量をみると、休日の休養方法に違いがあったことが分かる。この休養方法の違いによって、パフォーマンスが変化するという研究結果がある[(3)]。この研究は、本多准教授によって作成されたものであり、消極的休養と積極的休養では後者の方が、自覚疲労が減りパフォーマンスの向上がみられた。あくまで自覚疲労の軽減であるため、根本的な疲労回復はどちらも同程度であった。このような研究結果から、積極的休養の方が労働者のパフォーマンスを向上させることができるため、生産性も向上し所得の増加へとつながっているのではないかと考えている。また、積極的休養を行う人の多い都心部では SNS や web サイト等のきっかけによる出会いが多いため、簡単に仲間を集められる。したがって、連鎖的に積極的休養を行う人が増えていく傾向にあると考えられる。

### 5-2. 同調行動に伴う所得格差の拡大

IT 利用率が高い地域は都市であり、低い地域は地方であることがグラフから見て分かる。つまり、労働者のパフォーマンスでは、都市で高くなり、地方では低くなる傾向にある。このことから、都市と地方の格差が縮まることは考えにくく、格差が広がっていくことが考えられる。また、人間の行動特性として、同調行動[(4)]というものがある。同調行動とは、集団や他者が設定した標準や期待に沿った行動をとることである。つまり、自分が属するコミュニティによって何をすることが最適になるのかが変わってしまう。このように考えると、都市ほど積極的休養をする人が多くなり、地方では消極的休養をする人が多くなる。人が多くなれば、その休養方法がさらに定着され、同調行動がさらに多くなり、都市と地方の休養方法の差は拡大する。したがって、積極的休養によるパフォーマンス向上効果は都市にのみ起こり、労働者の生産性が向上し所得が向上する。このような同調行動に基づく変化が、現在の地方と都市の格差へとつながっていると考えている。

### 5-3. IT 利用率の変化による人材流出の懸念

インターネットを利用することで得られるメリットとして、地方と都市でも同様の情報を取れる。例えば、優秀な講師の講座や情報収集といった都市にいるからこそできるものが、オンライン講座やネット検索によって地方でも同じようにできるようになった。このメリットは地方にとって大変大きい恩恵だが、人材が都市に流れやすくなるというデメリットも発生していると考えている。情報が簡単に取れるため、自分を最大限に生かせる環境を探しやすいからである。つまり、IT 利用率が高くなればなるほど自分の最適な場所を求めて人材が移動する可能性が高くなる。情報格差をなくし、誰もが便利な社会になればと思いインターネットの普及を進めているが、地方にとっては逆効果の一面もあるということである。地方の活性化や魅力を向上させることよりも、IT 普及率を向上させてしまうと今後の人材流出はさらに悪化する可能性もあると考えている。また、人材流出についても同調行動が発生し、負の連鎖になることもあり得る。

### 5-4. まとめ

IT 利用率を起点に都市と地方の格差について検討した。IT 利用率を高め、多くの情報から好きな趣味を見つけ、趣味ができる環境が整えばパフォーマンスが向上する可能性がある。また、パフォーマンス向上に伴い生産性が向上し県民所得にも影響を及ぼすことも検討した。しかし、IT 利用率が上がれば人材の流動が活発になる可能性もあるため IT 利用率の増加がメリットだけではない。このようなトレードオフの関係の中で、何も優先すべきかを検討することが今後の課題になってくると考えている。

### 参考文献


（1） 劉　継生、“情報格差を解消するための対策に関する研究”、 通信教育部論集 第 21 号、P85-P102（2018 年 8 月）

（2） 総務省、“通信利用動向調査（世帯編)”、P-24、(2019 年)

（3） 本多　麻子、“積極的休養によるパフォーマンス向上と疲労回復効果”、日本心理学会、(2012 年 9 月)

（4） ケイン聡一、小池真由、中島健一郎、“同調行動研究のこれまでとこれから ―動機に着目する必要性―”、広島大学心理学研究 第 20 号、P121-P132、(2020 年)