

# データサイエンスリテラシー普及下の データ分析コンペティション開催で考慮したいこと

木川明彦<sup>†</sup>, 次郎丸沢<sup>‡</sup>, 鈴木治郎<sup>¶</sup>

<sup>†</sup> 立正大学, <sup>‡</sup> OME, <sup>¶</sup> 信州大学全学教育センター

<sup>†</sup> akihiko.kigawa@ris.ac.jp, <sup>‡</sup> ome@om-edu.jp, <sup>¶</sup> szkjiro@shinshu-u.ac.jp

キーワード オープンデータ, データ分析コンペティション, 情報検索

## 1 はじめに

文部科学省の AI 戦略 2019 が提唱されて 5 年が経とうとする現在, 国内の高等教育機関 (大学・高専・短大) においてはデータサイエンスリテラシー (以下 DS リテラシーと略す) を全学生が受講できる準備が進行中である。また 2024 年 3 月現在, 2 年目を終えようとする高校情報 I も, 構成内容自体は DS リテラシーに近いものをもつ。

AI 戦略 2019 ではリテラシーレベル以降の発展として応用基礎レベルを設けているが, そのカリキュラムが準備提供されるのは AI 戦略 2019 においてはリテラシーレベル受講者の半数に過ぎない。

本稿では主催者がオープンデータを用意した上で高校生の部および大学生・一般の部を設けて実施されている統計データ分析コンペティション (独立行政法人・統計センター主催) [1] に関する分析を合わせながら, DS リテラシー受講者が自身のデータ分析能力に挑戦できる機会としてデータ分析コンペティション (以下コンペと略す) をどのような形で設けるべきか, また, その意義について検討したい。

## 2 SSDSE 統計データ分析コンペティションについて

独立行政法人・統計センターでは SSDSE (教育用標準データセット: Standardized Statistical Data Set for Education) [2] を, データ分析のための汎用素材として, 作成・公開している。それは主要な公的統計を地域別に一覧できる表形式のデータセットである。

さらに高校生および大学生等を対象に, この SSDSE データセットを用いた統計データ分析の論文を募集し,

2018 年から統計データ分析コンペティションを毎年, 継続的に実施しており [1], その募集区分は「大学生・一般の部」と「高校生の部」に分かれている。

### 2.1 大学生・一般の部について

入賞内容を見る限り, 以下に説明するように, DS リテラシーレベルを修了しただけの大学生が参加しようと思えるものとは思えない。2023 年は審査員奨励賞を含めて全 10 件の受賞論文があり, その内, 大学生による受賞は 5 件あり, 以下の通りである。

入賞作品	データ分析方法
総務大臣賞 審査員奨励賞	小中学生の不登校率における環境要因分析 変数重要度に着目したクラスタリングによる 社会構造と睡眠時間の関係性の解析
審査員奨励賞	都道府県別のパネルデータを用いた合計特殊 出生率の決定要因
審査員奨励賞 審査員奨励賞	市区町村ごとの失業率の要因分析 子ども・子育て支援の充実は合計特殊出生率 を高めるか?

たとえば審査員奨励賞を受賞した大学生論文「市区町村ごとの失業率の要因分析」で<sup>\*1</sup>, どのようなデータ分析が行われているのか, 対話型生成 AI の一つ Gemini に「【この論文】<sup>\*2</sup>のデータ分析方法を説明してほしい」とプロンプトを与えて分析させると, 次のような回答を得た。

<sup>\*1</sup> 回答を全文引用できるよう, 生成 AI の回答が最短であった論文を選んだ。

<sup>\*2</sup> この部分に各論文のタイトルおよび受賞論文紹介サイトにある論文 URL を与えた。

### 審査員奨励賞におけるデータ分析方法

この論文では、市区町村ごとの失業率の要因を分析するために、統計的手法と機械学習手法が用いられています。

統計的手法としては、重回帰分析、Elastic Net 回帰分析、ランダムフォレスト回帰分析、主成分分析、k-means クラスタリングが用いられています。機械学習手法としては、Elastic Net 回帰分析、ランダムフォレスト回帰分析、k-means クラスタリングが用いられています。

分析の結果、労働力人口、事業所数、最低賃金、出生率、都市化が失業率に影響を与えることが示されました。

具体的には、労働力人口が多いほど、失業率は低くなる傾向があります。事業所数が多いほど、失業率は低くなる傾向があります。最低賃金が高いほど、失業率は高くなる傾向があります。出生率が高いほど、失業率は低くなる傾向があります。都市化が進んでいるほど、失業率は高くなる傾向があります。

なお、この論文では、分析対象としたデータは、2010年から2020年までの市区町村の人口統計データ、経済統計データ、社会統計データです。

筆者らは、これを含めた大学生の受賞論文全5件について生成AIを援用して分析方法に関する資料を作成した上で実際に論文内容を確認し、先に挙げた5件における分析方法について、順に次表の結果を得た。

データ分析方法
パネルデータ分析
ランダムフォレスト回帰分析他
固定効果モデル, ランダム効果モデル
ランダムフォレスト回帰分析他
DID法, 回帰分析

## 2.2 高校生の部について

同年の高校生受賞論文については、以下の通り、タイトルのみ紹介する。本稿の目的はDSリテラシー修了後の学生による応募を考察しているので、分析方法は論じない。

入賞作品	データ分析方法
総務大臣賞 優秀賞	生活の形態と女性の社会進出 大腸がん罹患要因の探究と罹患しにくい生活の提案
統計数理賞 統計活用奨励賞	時価に関する最適モデルの構築と手法提案 独自指標作成による地方創生の方法論と兵庫県活性化の提案
審査員奨励賞 審査員奨励賞 審査員奨励賞 審査員奨励賞 審査員奨励賞	日本人の英語能力の実態とその背景 都道府県ごとの学力の差 熱中症を防ごう！ 海水温からの降水量予測を目指して 日本の食料自給率を上げるために 兵庫県の宝「いかなご」を守る

## 3 コンペ参加者における分析行為の想定

SSDSE データセットは以下の6種類のデータから構成される [2].

名称	内容
市区町村	1741 市区町村 × 多分野 125 項目
県別推移	47 都道府県 × 12 年次 × 多分野 109 項目
家計消費	全国・47 都道府県庁所在市 × 家計消費 226 項目
社会生活	全国・47 都道府県庁所在市 × 男女別 × 社会生活 121 項目
基本素材	47 都道府県 × 多分野 90 項目
気候値	47 都道府県庁所在市 × 月・年 × 気象 42 項目

上記のコンペに応募しようという者は、おそらく何かを知りたいという動機のもとデータを集めて分析する、あるいは何かの機会に目にした、または自身で実施したデータ分析事例から疑問が生じて、その疑問を解決するためにデータを収集し、応募作品までに成長させるのではないだろうか。

このような応募者像を想定すると、上記 SSDSE のようなデータセットに、数々の分析方法を適用した結果、応募課題に到達するという見方はたぶん不適当だろう。このようなデータセットを用意しなくても（このデータの多くは総務省提供の e-stat などで入手可能である）、応募者自身が分析のためのデータ収集をしたいと考えたとき、その作業を支援することに重きを置く方が理にかなっているだろうと想像する。

さらにデータ分析方法について考えるとき、コンペでは上記の SSDSE 大学生・一般の部のように、DS リテラシーレベルからは想定しづらい高度な分析方法が使われていた。こうした分析方法を支援するのは、コンペ主催

者ではなく、応募者の周囲にいる指導者などであろう。

### 3.1 対話型生成 AI を利用するデータ取得

対話型生成 AI は 2022 年末に ChatGPT の登場以来、社会的に大きな反響をもたらしている。そうした一つ、Google 提供の Bard を利用して、世界各国の牛乳消費量を得るための質問文（プロンプト）を与えると、次の回答を得た（2023 年 12 月の実行）。

Bard の回答

FAO の牛乳消費量データの取得方法

1. **FAO ウェブサイトにアクセス** FAO のウェブサイト (<https://www.fao.org/faostat/>) にアクセスします。<sup>a</sup>
2. **検索バーで「牛乳消費量」を検索** 検索バーに「牛乳消費量」と入力し、検索します。
3. **データセットを選択** 検索結果から、以下のデータセットを選択します。タイトル: Milk - Production, Trade and Consumption データソース: FAOSTAT
4. **データ形式を選択** データ形式を選択します。CSV 形式と Excel 形式が利用可能です。
5. **データをダウンロード** 「Download」ボタンをクリックして、データをダウンロードします。
6. **データの利用** ダウンロードしたデータは、以下の目的で利用できます。レポートや論文の作成プレゼンテーションの作成データ分析その他
7. **データの注意点** データの定義や単位は、他のデータソースと異なる場合があります。データの品質は、FAOSTAT のデータ品質評価に基づいています。データを利用する前に、FAOSTAT の利用規約を確認してください。
8. **その他** FAO の牛乳消費量データに関する質問は、お気軽にお問い合わせください。

<sup>a</sup> 出力結果の保存のために Google ドキュメントに出力すると、URL 部分は「無効な URL を削除しました」に置き換わった。

このような回答を得るのが容易であれば、データ分析で何かを明らかにしたいときのデータ収集は、生成 AI の利用できわめて容易になるだろう。ところで、この回

答事例の FAOSTAT サイトは英語で書かれている。現在ではブラウザ上の日本語翻訳機能が利用できるのですが、英文サイトから上の回答に示す項目を見つけるのも困難ではない。

さて肝心の、この回答を得るための質問文（プロンプト）であるが、生成 AI の癖を活かすための特別な工夫はしていない。次がそのプロンプトである。

データ取得のための質問

世界各国の牛乳消費量に関するオープンデータを取得したい

なお 2024 年 2 月現在、Google 提供の対話型生成 AI は、Bard から Gemini に交代した。そのためか同じプロンプトに対しては、データ提供サイト（OECD 等いくつかの候補）を最初に示した。その回答にもとづき、プロンプトにある「オープンデータ」を「FAO 提供のオープンデータ」に修正することで、ほぼ同じ回答が得られた [3]。生成 AI への指示回数が 1 回増えたとはいえ、データ取得作業は、質問文の考案を含めて容易であることには違いないだろう。

## 4 私たちの提案

以上の分析のもとに、私たちはオープンデータ利用のデータ分析コンペティションを、あらかじめデータセットを用意せずに、代わりに生成 AI を活用するオープンデータ検索支援の案内をすること [3] を提案したい。

このように考えた上で、私たちの提案するデータ分析コンペティションの全体像は以下の通りを考えている。

- DS リテラシー修了者以上を想定した案内を行う
  - 本学会の全国大会などの発表状況を考慮して、コンペに取り組みやすい分野（環境など）を設けるか検討する
- 発表手法はポスターセッションに近いものとする
  - ポスターの発表 1 件ごとに聴衆全員が回っていけるようにする
- 利用データはオープンデータ利用を推奨する
- 募集においては利用データ自身（あるいはリンク）、分析手法（参考文献あるいはリンク）、分析結果を含むアブストラクトを提出
  - PDF ファイルを提出してもらえば運営側で A0 版に印刷用意する

- 持参する場合は A0 版とする
- 論文募集とは異なり，論文集の出版はしない
- コンペと学会ポスターセッションとの違いなど想定質問集を用意する
- 評価基準
  - 図表やグラフなど視覚要素にどれだけ訴えているか
  - 社会的意義
  - 分析方法の的確さあるいは独自性
  - プレゼンテーションのわかりやすさ

他のデータ分析コンペティションの例には，ポスターセッションに近いものに，高校生以下を参加対象とする全国統計グラフコンクール [4] があり，また経済産業省の RESAS 利用を前提としたものに地方創生☆政策アイデアコンテスト [5] などがある．これらのコンペも，私たちが実施計画を進めているコンペの運営方法の参考にした．

## 5 おわりに

データ分析コンペティションを開催するにあたり対話型生成 AI を用いた支援を行えば，オープンデータの利用はさらに容易になり，私たちが本学会で企画するデータ分析コンペティションにおいて学生たちが参加する助けになると考えている．

## 参考文献

- [1] 独立行政法人・統計センター，教育用標準データセット，<https://www.nstac.go.jp/use/literacy/ssdse/>
- [2] 独立行政法人・統計センター，統計データ分析コンペティション，<https://www.nstac.go.jp/use/literacy/statcompe/>
- [3] 鈴木治郎，『オープンデータ取得のための生成 AI 利用の比較検討』，第 6 回国際 ICT 利用研究学会・第 15 回研究会，2024 年
- [4] 公益財団法人・統計情報研究開発センター，統計グラフ全国コンクール，<https://www.sinfonica.or.jp/tokei/graph/>
- [5] 内閣府地方創生推進室，統計グラフ全国コンクール，<https://contest.resas-portal.go.jp/2023/>