特集にあたって

生田目　崇（中央大学）

本号は，令和最初の「データ解析コンペティション」の研究成果に関する特集である．令和元年度のコンペティションでは都内タクシー（特別区・武蔵野・三鷹交通圏）のプローブデータを提供した．

本コンペティションでは，93 チーム延べ 700 名の参加を得て開催された．例年 3 月に開催している成果報告会は，他例に漏れず新型コロナウイルスの影響を受け，例年より遅れて 6 月にオンラインで開催せざるを得なかったが，大変充実した成果を見ることができた．本特集は，参加チームから査読付き論文への投稿として募集し，8 編の投稿を得て，査読の結果 3 編を採択し本号に掲載いただいた．

本コンペティションのデータは，みずほ情報総研株式会社のご協力により提供された．プローブは計測の世界では計測器の端子や電極を指す単語として広く知られているが，本データのような道路交通におけるプローブデータとは，車両を一つのセンサーとして各車両に関する位置や速度，その他の状況などを観測し集めたデータを指す．データはセンサーや測定機により自動的に収拾されるものであり，こうした仕組みはある意味で業界全体で取り入れられた IoT システムということも言えよう．本データには，車両 ID，車両データ（車色など），運転中のドライバー ID，緯度経度，車両状況（空車か実車かなど），方向，速度といったデータが含まれる．データ期間は 2016 年から 2018 年の 2 年間であり，対象となるタクシーはおよそ 7,000 台に上る．各車両のデータについては，エンジンがかかっている状態であれば，一定の時間間隔で収集されているが，従来のタクシー無線か公衆回線を用いたものかによってその時間間隔は数秒から 30 秒程度まで異なる．本データは，変数の数はさほどでもないが，数千台のタクシーについて上記のように，自動的にデータを取得するため，データ量は過去最大の 600 Gbyte に達した．過去最大のデータであり，参加された方々は前処理に相当苦労されたものと思われる．そうした前処理をしながら，各チームで分析目的やどのように分析を実施していくかを考え，実行していく過程にはいくつもの苦労があったことは想像に難くない．投稿された論文においては，各種の創意工夫や最新の分析手法の適用など野心的なものも多く，この場でそれらすべてを紹介できればよかったが，査読の結果，採録に至らないものもありその点は残念であった．

本コンペティション開催にあたり，大変貴重なデータを提供いただいたみずほ情報総研株式会社には改めて謝意を表します．提供するデータがなくては開催できないコンペティションであり，趣旨をご理解いただいたことに重ねての謝意を表します．また，分析環境として希望チームに分析ツールを貸与いただいた株式会社 NTT データ数理システム，さらに，短期間に貴重なコメントをいただいた査読者の皆様にも感謝を申し上げます．

コンペティションでは，学部生主体のチームも多く参加され，データハンドリングで根を上げるのではないかという危惧もあったが，一チームも欠けることなく各研究部会で最終報告をいただいたことは大変うれしい．こうしたコンペティションで本データのような構造化データにおいてこれ以上のデータはしばらくはないと思っているが，過去のコンペティション開催を通じて，提供データにどのような変遷があったのかなどをまとめたものを本誌の昨年 11 月号に寄稿したので，ご興味ある方はご一読いただきたい．こうしたデータを扱う知識や技術が一般化し，あらたな価値につながる結果を出せるようになったことは大変喜ばしい．

さて，今年度の本コンペティションは大規模アンケート調査データを提供し開催している．新型コロナウイルスによりさまざまな影響が続いている昨今ではあるが，提供データには新型コロナウイルスが日本社会に影響を与えた後の期間のアンケートデータも含まれており，生活者の声の一端を分析できるのではないかと思われる．本記事を書いている 2020 年 11 月現在においては，今年度もオンライン開催になる可能性も高いと思われるが，ご興味ある方はぜひ成果報告会にご参加いただければ幸いである．