まえがき

本書は，R やデータ分析の初歩を学んだ読者を対象として，データ分析を実行するプロセスで必要となる知識や方法について習得するための書籍である．

通常，データ分析といえば，多変量解析（重回帰分析，判別分析，主成分分析，クラスター分析等），機械学習（決定木，サポートベクタマシン，ランダムフォレスト等），時系列解析（AR モデル，ARMA モデル，ARIMA モデル，ARCH モデル，GARCH モデル等）などの手法が上げられることが多い．しかし，実際にデータを分析して有用な知見を得るためには，適切な目標に基づいて分析計画を立案したうえで，データを収集・蓄積し，適宜，加工や変換を行った後に分析手法を適切に適用する必要がある．

本書では，このようなデータ分析プロセスを実現できるようになることを目指して，収集・蓄積したデータに対して加工や変換を行い，データから相関やパターンなどの知見を抽出するための基本的な考え方や処理について，R の実装方法を交えて説明する．統計解析や機械学習の分析手法を R で実行する方法についてはすでに良書が多数出版されているため，本書では著者の経験に基づいて，機械学習のアルゴリズムのチューニング方法やクラスのデータ数に偏りのあるデータへの対処など，実際にデータを分析するにあたって必要となる事項について重点的に扱う方針とする．

本書の最後には，実際のデータ分析の例を取り上げる．データ分析はどのような手順で進んでいくのか，知見を発見するためにどのような分析を行うのかについて，R での実行方法も含めて説明する．公開されているデータを用いるため，現場の生々しいデータ分析からは少し遠いように感じられるかもしれないが，それでもデータ分析の実際について理解を深めていただけるのではないかと思っている．

本書の構成は以下のとおりである．

第 2 章では，データ分析プロセス全般にわたり必要になる R のデータ操作について説明する．`dplyr` パッケージや `tidyr` パッケージ，`readr` パッケージなど，比較的新しい話題についても取り上げる．

第 3 章では，分析手法が適用できるようにデータを加工・変換する処理について取り上げる．欠

損値の対処，外れ値の検出，連続値の離散化，属性選択について説明する．これらの処理はデータのクレンジングとも呼ばれ，データ分析の 8 割から 9 割を占めるともいわれており，非常に重要である．

第 4 章では，加工・変換済みのデータから知見を発見するための方法論について説明する．予測モデルの構築とその評価，頻出するパターンの抽出について取り上げる．特に予測モデルについては，どのようなプロセスで予測する問題を設定し，モデルを構築していくかについて，著者の経験を踏まえて重点的に説明する．

第 5 章では，データ分析プロセスの例を取り上げる．実際のデータに対して，本書のこれまでの章で説明した手法を用いた分析の例について説明する．

また，本書では R を用いたデータ分析のプロセスについて説明を行うが，特に前処理やデータの変換についてはデータの規模が大きいほど R は最適なツールとはいえなくなる．そこで，本書では R の実装例を示すが，サポートページでは Python のプログラムも提供する．Python は汎用的なスクリプト言語であり，近年注目を集めている．

本書の執筆にあたり，非常に多くの方々のご協力，ご支援をいただいた．同志社大学の金 明哲氏には執筆の機会をいただき，ことあるごとに原稿に目を通して有益なアドバイスをいただいた．共立出版編集部の横田穂波氏には，なかなか筆が進まない状況にも辛抱強く便宜を図っていただいた．また，より良い内容を目指して直前まで加筆修正を繰り返す著者に対しても，忍耐強くご尽力いただいた．横田氏の並々ならぬご理解とご尽力がなければ，本書が世に出ることは決してなかったと断言できる．

本書では，可能な範囲で実データを使用した分析例を例示しようと心がけたつもりである．Craig K.Enders 氏，ならびに Guilford Press 社の Mandy Sparber 氏，C.Deborah Laughton 氏には 3.2 節で欠損値への対応について理解するために用いる従業員の知能指数と業務成果の関係を表すデータの使用および配布を許諾いただいた．Clopinet 社の Isabelle Guyon 氏と Orange 社の Vincent Lemaire 氏には，4.1.11 項で使用する KDD Cup 2009 のデータセットの使用を許諾いただいた．Chun-Nan Hsu 氏には，4.1.12 項，4.2.2 項，4.2.4 項で使用する食料品店の POS データ (Point of Sales) である `Tafeng` データセットの使用を許諾いただいた．Andrew T.Campbell 氏をはじめとするダートマス大学の StudentLife Study プロジェクトの方々には，第 5 章で使用する `StudentLife` データセットと関連する文献の図の使用を許諾いただいた．

さらに，株式会社金融エンジニアリング・グループの黒柳敬一氏，DATUM STUDIO 株式会社の里 洋平氏，市川太祐氏には執筆中に有益なコメント，アドバイスをいただいた．以上の方々に深くお礼を申し上げる．また，本来であれば，より多くの識者に完成した原稿を見ていただく予定であったが，著者の時間管理の問題に帰して，その機会を逸してしまったことは非常に残念であったことを付記しておく．

そして，本書の完成を見ることなく 2014 年 12 月に他界した父・哲弘と，いつも温かく見守って

くれる母・延子に特に本書を捧げたい.

本書が読者のデータ分析に対する理解を深め，日々の分析において少しでも役に立つならば，著者としてこれに勝る喜びはない.

2015 年 5 月

福島　真太朗