本シリーズの編集にあたって

社会の進化に伴い，統計科学の環境が大きく変化している．その主な変化として次のような点があげられる．1) データの収集の方法が多様化されている．2) データの平均サイズがますます大きくなっている．3) データの流通が容易になっている．4) 統計計算やシミュレーションに必要となるコンピュータがますます安価になっている．5) 統計計算やシミュレーションの専用ソフトが無料で入手可能になった．6) 統計科学の役割の重要性の認知度が向上している．

このようなさまざまな変化は，統計的データ解析の新しい手法の開発と応用を促し，データマイニング (data mining) や統計的機械学習 (statistical machine learning) のような新しい研究分野が生まれるようになり，その応用が急速に広がっている．従来の統計学，近年のデータマイニングや機械学習（マシンラーニング）に関する定義はいろいろあるが，共通点はデータを対象としていることであるので，本シリーズではこれらを包含する用語として，狭義のデータサイエンス (data science) を用いることにする．

データサイエンスは，広義ではデータの収集，加工，蓄積，管理，流通，解析，マイニングなど，データの流れの上流から下流までを貫く科学である．昨今，データサイエンスは，工学，医学，薬学，生命科学，社会科学（社会，経済，マーケティングなど），心理学，教育学はもちろんのこと，文化学のような，従来は統計学やデータ解析があまり応用されていなかった分野でも，データサイエンスの手法による斬新な研究成果が多く報告されている．データサイエンスは，あらゆる分野において必要となる万人の科学と言っても過言ではない．

データ解析の手法のほとんどは数理的理論に基づいて開発されているので，データサイエンスに関する解説書では，数式を避けると厳密な説明ができなくなる．非理工系の研究者の中には数式が苦手である方が多いため，非理工系の研究分野におけるデータサイエンスの適用が遅れている．一方，データ解析のツールを用いると，数理的な理論が分からなくても，データを入力すると何らかの結果が出力され，形式上はデータ解析が可能な時代になっている．しかし，データ解析の理論に関する理解が不十分であると，統計手法の利用を間違えたり，出力された結果の解析を誤ったりする可能性がある．

データ解析を行うには，用いる手法の数理的理論の理解だけではなく，ツールを用いてデータを解析しなければならない．そのためには，データサイエンスの基礎理論を理解した上でツールを用いてデータを操作し，データ解析やデータマイニングを行うことが望ましい．データ解析やデータマイニングの手法は，データの構造と目的に依存する．万能なデータ解析やマイニングの

手法はない．データ解析やマイニングを行う際には，データの構造や目的に合う手法を用いることが必要である．そのためには，用いる手法の理論を正しく理解することが必要である．

データ解析の手軽なソフトとしては，表計算ソフト Excel や Calc がある．前者はマイクロソフト社の有料ソフトであり，後者はサン・マイクロシステムズ社が開発したフリーソフトである．最近，個人，法人を問わず，ほとんどのパソコンには Excel がインストールされていることもあり，広く利用されている．表計算ソフトは，データの整理や簡単な計算には便利なツールであるが，高度なデータ解析を行うためには，プログラムを作成するか追加ソフトを用いることが必要である．また，これらのソフトは列の数に制限があり，大量のデータ解析には向いていない．その一方，データ解析の専用ソフトとしては SAS，SPSS，S-PLUS などがあるが，これらは高価であるため，個人のポケットマネーでは購入しがたく，恵まれている環境でなければ使用できない．

このようなことから，1990 年代にニュージーランドのオークランド大学統計学科の Ross Ihaka とアメリカのハーバード大学の Robert Gentleman により R（R 環境，R 言語とも呼ぶ）というデータ解析ツールの開発が始められ，1997 年からは多くの賛同者が加わり，オープンソース方式で開発が続けられている．R はフリーソフトであり，インターネットが接続された環境であれば，誰でもどこでも自由にダウンロードできる．R は，基本的な統計計算の環境と専用パッケージの利用環境を提供している．2009 年の現在，公開された R 専用のフリーパッケージの数は 2 千を超えている．R では，表形式のように定型化されたデータ処理やモデリング，データマイニング，機械学習などはもちろん，定型化されていない遺伝子情報のデータ，画像データ，音声・音楽データ，テキストデータなどを解析することも可能である．R は，高度なデータ解析，繊細多様なグラフィックスの作成，データマイニング，機械学習，シミュレーションなどを行うツールであり，伝統的な統計計算やデータ解析の概念を超えたツールとして発展し続けている．従来は，研究者が考案したデータ解析の方法をエンドユーザが使用するまでには長い時間を要したが，R の普及のおかげで，研究者が考案した新しい方法をエンドユーザが使用できるようになるまでのサイクルが大幅に短縮されている．

R による統計学に関する単行本がわが国で初めて刊行されたのは 2003 年である．約 5 年の間に R に関する訳書・和書の数はすでに 30 冊を超えるようになった．これが R 普及の勢いを物語っている．しかし，その中には R による初級統計学や R のマニュアル形態のものが多く，高度なデータ解析やデータマイニングに関する理論を系統的に説明し，その方法を R で実践する，いわゆる理論と実践を両立したものが少ない．

そこで，数理的な基礎が一定程度ある方は，関連手法の数理的理論を理解し，R による実践を通じてその方法の理論と応用を学び，数理的基礎が弱い方々は，R を用いて実践的に入門し，数理的理論を徐々に理解するようにと，数理に強い，弱いに関係なく幅広く使用できる本を提供することが本シリーズの主な目的である．ただし，企画した時点ですでに上記の理念と一致する本が刊行されている分野もある．それらの内容に関しては，重複を避けている．本シリーズでは，可能であれば社会的ニーズに応じて新たな内容の巻を追加していく予定である．

各巻の著者は，それぞれの分野で教育と研究にご活躍されている専門家である．ご多忙にもかかわらずご執筆をお引き受けいただいたことに感謝する．

本シリーズがデータサイエンスの発展に少しでも寄与できれば幸いである．

編者　金 明哲

まえがき

位置によって変量または変質する地球上の現象に関心をもつとき，その現象を解釈するためには，その現象から得られるデータ（地理空間データ）の分析やモデリングが必要になる．

地理学や地質学において位置への関心が高いのは当然であろう．しかし，今日では，位置情報が含まれるデータを扱うあらゆる分野で，地理空間データの分析やモデリングへの関心が高まってきている．オープンソースソフトウェアの R は優れた統計解析環境であり，空間統計解析を得意とするが，残念ながら R による空間統計解析または地理空間データ分析を解説する和書は存在しなかった．また，地理空間データ分析の和書はいくつか刊行されているが，理論を述べるにとどまり実践的な書籍はなかった．本書は，既刊の類書にはない，理論と実践の両方を満たす書籍であると考えている．

地理空間データ分析は，空間統計解析/分析，空間解析/分析，空間データ解析/分析とも呼ばれるが，それは分野の歴史的背景や中心となる技法が異なるためである．地理空間データを扱う分析理論が学際的に統合されつつある昨今では，地理空間データ分析と総称するのもよいと私は考えている．本書では，特に使い分けが必要な場合を除いて「地理空間データ分析」で統一することにする．

本書は，地理空間データに関する基礎理論と分析手法について述べた解説書である．理論的背景の解説に加えて，R による事例解説により，読者が自分自身の手で実践的に理論・方法・技術を確認できるように構成されている．本書の対象は，地球統計学，計量地理学，地理空間情報学を目指す学生のみならず，位置情報が関与するさまざまな分野，環境学，疫学・公衆衛生学，犯罪科学，経済学，生態学，都市計画学，マーケティング，その他の実務家，専門家，学生を幅広く対象としている．

本書の構成は，地理空間データ分析の基礎理論と R による実践方法を解説した主要部分と，補足的な R の操作方法や関連知識を紹介した付録部分とに分かれている．第 1 章では，地理空間データの特性やデータ構造モデルの種類，さらに地理空間データ分析に必要な地理学的基礎知識を述べる．第 2 章では，カーネル密度推定法・クリギング・経験的ベイズ推定法を含むさまざまな視覚化の理論および実践方法について解説する．第 3 章では，地理空間データの分布パターンを調べるための基礎的理論について解説し，特に地域集積性の検出方法について総括する．第 4 章では道路など交通ネットワークにおける分析について述べ，第 5 章では，地域要因を説明変数とした地理空間相関分析についてこれまで開発された代表的なモデルを紹介する．第 6 章では立地分

析の簡単な紹介を行っている．付録として，空間集計や空間計測など地理空間データ操作の解説なども収録している．さらに，**sp** パッケージや **rgdal** パッケージなど基盤となるパッケージが提供するクラスやメソッドも付録として加えている．

紙面の都合から，地理空間データ分析を網羅的に扱うことができず，内容を取捨選択し，大幅な割愛や省略を加えた．そのため，十分に説明できていない単元も多く，物足りなく感じる読者もいるかもしれない．そのような読者は関連洋書をあたられたい．また，本書ではベイズ法をごく一部を残して全面的に割愛した．GRASS や Quantum GIS など GIS ソフトウェアとの連携に関する解説も全面的に割愛している．付録 A でそれらを扱うパッケージを紹介しているので，興味のある読者はそれらのヘルプをあたられたい．

本書の執筆にいたる研究活動において，実に多くの人々から直接・間接的な助言や指導を受けた．GIS や地理空間データ分析のイロハを教授していただいた立命館大学の中谷友樹准教授，Health GIS の武者修行のために渡米した私を暖かく迎え入れていただいた Centers for Disease Control and Prevention (CDC) の Allen Hightower 博士，マンツーマンで **sp** パッケージや空間データ操作をデモンストレーションしてくださったノルウェー経済大学の Roger Bivand 教授，その他にもたくさんの諸先輩や同僚諸氏から影響を受けた．心より感謝申し上げたい．また，R の SIG (special interest group) コミュニティーである R-sig-Geo メーリングリストや筑波大学の岡田昌史講師が運営する Rjpwiki からも貴重な情報を得ることがたびたびあった．

本書では，パッケージ名索引を見れば分かるとおり，たくさんのパッケージを利用している．特に本書の根幹をなすパッケージ (**sp, maptools, splancs, DCluster, spdep**) の作者であるユトレヒト大学の Edzer Pebesma 教授，前出の Roger Bivand 教授，Genentech 社の Nicholas J. Lewin-Koh 博士，ランカスター大学の Barry Rowlingson 博士および Peter Diggle 教授，カステラ・ラ・マンチャ大学の Virgilio Gómez-Rubio 助教，バレンシア大学の Juan Ferrándiz-Ferragud 教授および Antonio López-Quílez 教授に感謝したい．少なくともこれらのパッケージがなければ，本書は存在しなかった．

編集者である同志社大学の金明哲教授から執筆の誘いがあってから実に 2 年以上が経過してしまった．我ながら呆れる遅筆を反省する．執筆の間にも R をめぐる地理空間データ解析環境は進歩し続け，執筆中に古くなった内容を捨てて何度か新たに書き直した．R が進歩する以上，基礎理論はともかく，実践的な部分で最新の R に合わない部分が将来に出てくるかもしれない．何らかの対処ができればよいと思っているが，それが叶わなかったときはご容赦願いたい．

最後に，編集者の金明哲教授および共立出版（株）の横田穂波氏・國井和郎氏には本当に辛抱強く原稿の完成を待っていただいた．両者には心から感謝申し上げたい．

2010 年 6 月

谷村　晋