報告
Reports

正しい分析結果を導くためのデータ前処理

―分析者に求められる正確なデータ研磨のスキル―

菊川康彬 | （株）帝国データバンク

一般

データ分析における前処理

前処理工程の重要性

ビッグデータ時代と呼ばれて久しく，従来と比してデータ分析者が扱うデータのボリュームは増大し，データの構造もより複雑になっている．アメリカの調査会社である IDC（International Data Corporation）の調査によると，世界的なデータ量は 2010 年から 2020 年までの 10 年間で，約 40 倍にものぼる 40ZB（ゼタバイト）まで膨れ上がる見込みである．ゼタバイトとは，テラバイトの 10 億倍に相当する単位で，40 ゼタバイトがどれだけ膨大なデータ量であるかが分かる．さらに，扱うデータ量の増大やデータ構造の複雑化だけでなく，データの取得が容易になったことで複数のデータを組み合わせて分析することが可能になるなど，分析の幅が広がっているという潮流もある．そのような状況下でデータ分析を行う上では，分析の全工程のおよそ 8 割を占めるといわれている「データ研磨」，いわゆるデータ分析の前処理に該当する作業が非常に重要となる．なぜならば，データの形式が異なるデータセットを組み合わせる際にはデータ構造の統一が必要だからである．異種のデータを組み合わせる場合にデータ構造の統一が必要なことはいうまでもないが，同種のデータの場合でも時系列での比較を目的としたときにデータ構造の統一化の必要が生じ得る．

また，ビッグデータ時代においては，利用目的が明確ではないデータも含めて取得が容易になったという特徴もある．従来は利用目的が明確である上で収集されたデータを扱うことが大半であったが，現在は「利用目的は不明確だが収集・蓄積されたデータ」までも利用可能であるケースが増加している．特に，企業の内部で蓄積されたデータに代表されるように，「従来利用してこなかったが，データは蓄積されている」というデータも分析の対象となってきている．利用目的が定まっていないデータについては，データ分析者にとって分析しやすいデータ構造ではないことが往々にしてある．このように，時代の変化に伴い分析者が扱うデータの量や構造も変化している中で，データの前処理工程に割かなければいけない時間も増大しているといえる．

データ研磨の定義

本稿では，「データを分析が可能な形式にするための前処理」をデータ研磨と定義する．データ研磨は，データ分析をするための前処理に特化している点が特徴である．なお，生データに対して前処理を施すことでデータ分析が可能な状態にすることを目的としているため，データから価値を取り出すことができる形まで生データを磨くという意味から「研磨」と命名した．そのため，国立情報学研究所の宇野毅明教授が提案している精度向上を目的とし

てデータを再構成するデータ研磨とは異なるものである点に留意されたい[☆1]．また，データウェアハウスを構築するためのETL（Extract/Transform/Load）工程と処理工程そのものは類似しているが，データ研磨は構築するデータの最終形が「データ分析が可能な形」としている点に違いがある．

一般的に，データの前処理に関連した用語として，データクレンジング，データクリーニング，データラングリング，データ加工，データ整備など種々な呼称がある．しかし，用語の使われ方についてはコンセンサスがない状況である．それらとデータ研磨の違いを明らかにするために各用語の特徴を整理したい．前述した用語の中ではデータラングリング，データ加工が最も広義な用語でデータ前処理全般を表すのに対し，データクレンジング，データクリーニングはデータの表記揺れの統一やコード体系の統一，データ誤入力の修正，不要レコードの削除など「clean」に由来するようにデータを「キレイにする」部分に特化して使われることが多いのではないだろうか．データ整備に関しては，クレンジング，クリーニングを包含し，さらにはデータ形式の標準化など構造を整備する意味合いが強い（クレンジング・クリーニングと同義で扱われることもあれば，前処理全体という意味で使用されることもある）．それ以外にも，機械学習の分野においてはデータを加工することで新たな特徴量を作成する特徴量エンジニアリングという手法があるが，データの前処理というよりはデータ分析の前段階で発生する工程といえる．

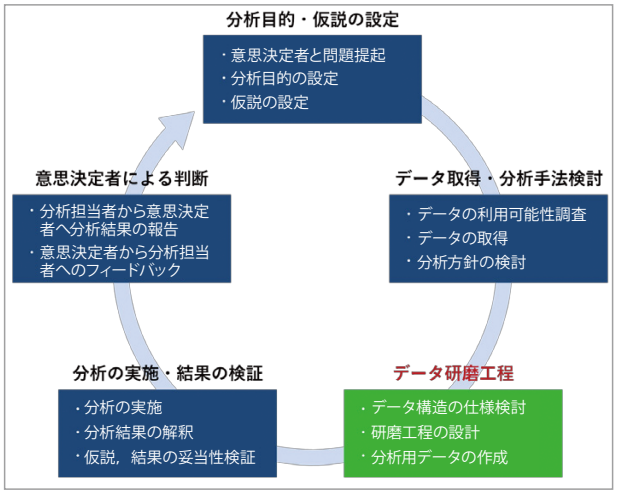
データ研磨を正確に行うことは分析時に使用するデータの品質を担保するためにきわめて重要である．データの品質が損なわれるとデータの分析結果にも影響が出てしまう．しかし，データ研磨にかかるスキルを体系的に学ぶ機会は少ない．多くはデータ分析者の自助努力による習得である．そのため分析者によってデータ研磨スキルの習熟度には大きなばらつきがある．データ分析の8割を占めるとされるデータ前処理は，データ研磨スキルを身につけることで効率化が可能である．そして効率化ができると，より比重を置くべき分析工程に時間を割くことができるため，データ分析者は分析手法だけでなく研磨スキルについても習得が求められる．

EBPMとデータ研磨

EBPM（Evidence-Based Policy Making：証拠に基づく政策立案）が叫ばれる昨今，データに基づいた意思決定が重視されている．データ分析を行い，データに基づいた意思決定を行うために必要なサイクルを，帝国データバンクは次のように設定した（図-1）．

1. 意思決定者や顧客との分析目的・仮説の設定
2. 分析に必要なデータの取得と分析手法の検討
3. 分析用データ作成のためのデータ研磨
4. 分析の実施と分析結果や仮説の検証
5. 意思決定者による判断

このようなサイクルをまわしていくことにより，ようやく活用できるレベルのデータ分析となっていく．当然，サイクルをまわしていくためには，データ研磨の工程も複数回通ることになる．データ研磨のスキルを身につけることは，試行錯誤のスピードを上げることにもつながっていく．そして，いかに



■図-1　EBPMのために必要なサイクル

☆1 http://research.nii.ac.jp/~uno/CREST/particlization/particlization.html

分析目的が明確で，分析に必要なデータを取得でき，様々な分析手法を身につけたとしても，分析に必要なデータ研磨を自ら行うことができなければ分析にすら辿りつくことはできない．前処理ができないことが分析全体のボトルネックになることは容易に発生し得る．また，データの扱いを誤ると分析結果も誤りとなってしまう．正しい分析結果を導き出すためにはデータの前処理部分でデータを適切に扱うことが大前提となる．

データ研磨が必要な事例

次に，どのような場面でデータ研磨が必要になるのかを具体的な統計を例に説明する．本稿では，事例の分かりやすさのために，オープンデータであり一般性の高い公的統計を例として挙げる．補足であるが，あくまで事例として公的統計を対象に説明するものであり，近年の e-Stat（政府統計の総合窓口）では csv ファイルや Excel ファイルでの公開だけでなく，DB（データベース）や API の機能も徐々に拡張されており，後述する研磨をせずに整備されたデータを取得することが可能なデータも存在することには留意されたい．

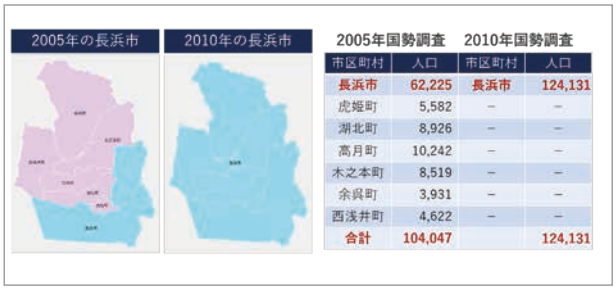
市区町村の統廃合

市区町村単位で集計されたデータに関しては，「市区町村の統廃合」を考慮する必要がある．2000 年 1 月 1 日時点では 3,235 あった市区町村も，2019 年 1 月 1 日時点では 1,724 まで減少（統合・合併）している．2000 年代前半は特に「平成の大合併」と呼ばれる大規模な動きがあったため，時系列データで 2000 年代前半のデータも使用する際には特に注意が必要である．このように，市区町村の統廃合情報の反映は，たとえば時系列でデータを比較する際に「同一の地域」において比較可能である状態を作るために行うことがある．市区町村の統廃合を考慮しない場合どのような影響が生じるかを，滋賀県長浜市を例に紹介する．使用するのは，2005 年と 2010 年の国勢調査のデータである．

滋賀県長浜市は，2010 年 1 月より，6 つの町を吸収して新たな長浜市となった（図 -2）．国勢調査のデータを取得すると，2005 年の長浜市の人口は 62,225 人，2010 年の長浜市の人口は 124,131 人となっている．このデータを市区町村の統廃合を考慮せずに時系列的に使用した場合，2010 年の人口は 2005 年比で 1.99 倍という数値になってしまう．無論この値は誤りであり，2005 年当時は長浜市に吸収されていなかった 6 つの町の人口を考慮した上での比較をしなければならない．市区町村統廃合を考慮すれば，2005 年は 104,047 人，2010 年は 124,131 人となり，2005 年比で人口は 1.19 倍になっていることが分かる．このように比較対象を統一することで，正確な時系列での比較が可能となる．余談ではあるが，稀なケースとして越境合併と呼ばれる合併がある．2005 年 2 月 13 日より，長野県木曽郡山口村は岐阜県中津川市へ編入となった．このように都道府県の境界をまたいだ市区町村の合併のことを越境合併という．市区町村データを足し上げて都道府県の合計値を算出する際などには注意が必要である．

総務省「住民基本台帳人口移動報告」

住民基本台帳人口移動報告では，人の社会的な移動，つまりは転入・転出をとらえることができる．From-to データとも呼ばれ，人口移動を把握する上では重要なデータとなる．オープンデータで誰でも利用が可能なデータではあるものの，データの公表



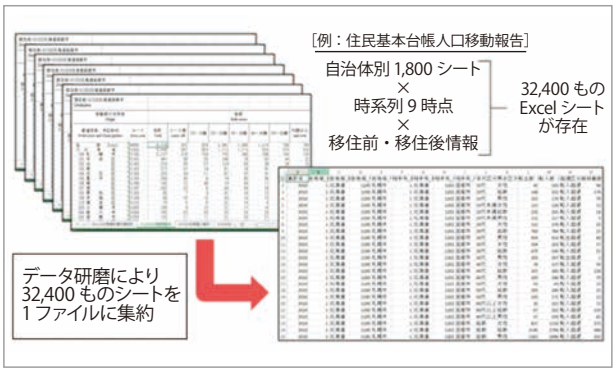
2005年国勢調査		2010年国勢調査	
市区町村	人口	市区町村	人口
長浜市	62,225	長浜市	124,131
虎姫町	5,582	–	–
湖北町	8,926	–	–
高月町	10,242	–	–
木之本町	8,519	–	–
余呉町	3,931	–	–
西浅井町	4,622	–	–
合計	104,047		124,131

■図 -2　滋賀県長浜市を中心とした市区町村統廃合

形式が少々複雑である．まず，都道府県別に Excel ファイルが作成されている（47 個の Excel ファイル）．さらに，市区町村間での人口移動も把握できるデータであるため，各都道府県の Excel ファイルの中に，市区町村の数だけシートが分かれている．都道府県合計や政令市合計，政令市区部の情報も取得できるため，各年で自治体の数に相当する約 1,800 の Excel のシートに分割されていることになる．その上，詳細は割愛するが「移動前の住所地」と「移動後の住所地」という 2 つの軸で作成されたデータが存在するので，各年に対して約 3,600 にのぼる Excel のシートが存在する．たとえば 2010 年～2018 年までの 9 時点の時系列データで分析を行おうとした場合，3,600 × 9 で約 32,400 もの Excel のシートに分割して格納されたデータを統合するところから始めなければならない（図 -3）．もちろん，統合が完了した後には市区町村統廃合の対応も必要となる．数シートあるいは数十シート程度であれば力ずくで対応することも可能かもしれないが分割された大量のデータを扱えるデータ研磨スキルを習得していれば，一括で対応することが可能となる．

厚生労働省「衛生行政報告例」

データ研磨では，単純な処理だけではなく使用するデータの特性を丁寧に理解することも重視している．同一の統計調査で時系列的に公表されているデータも，必ずしもすべてが同一のフォーマットで作られているわけではない．衛生行政報告例の隔年報は，2 年おきに病院数や診療所数などを把握できる統計である．しかし，調査時点によってデータの区分（病院数や診療所数に関する内数の有無）やデータの開始位置，都道府県コードの有無等が異なる．同じ項目のデータでも，入力されているデータの位置が年によって変わり，すべての時系列データを同一フォーマットで読み込むことができない．まずは各時点でバラバラに格納されたデータを同一のものと定義することにより，正確な時系列データの構築が可能となる．



■図 -3　住民基本台帳人口移動報告の研磨イメージ

厚生労働省「職業安定業務統計」

職業安定所別に集計された有効求人数，有効求職者数，有効求人倍率のデータが取得できる統計である．職業分類別にデータが存在するものの，職業大分類・職業中分類含めてすべて，Excel のシート別に並列で格納されている．そのため，まずは Excel のシート名を参照し，シート名をデータとして格納することが，一般的に行われる方法だと考えられる．しかし，厚生労働省が公表している形式では Excel 上で視認性が高くなるように，中分類は大分類よりも下位のカテゴリであることを示そうと，シート名の先頭に半角スペースが入力されている．そのようなデータの場合，たとえば統計解析ソフトの SAS 等を用いてデータを自動的にインポートしようとしてもスペースが先頭に含まれていることでエラーが発生してしまう．人間にとって見やすいデータ形式と機械処理しやすいデータ形式は異なるため，データ研磨を通じて生データを分析可能なデータ形式にする必要がある．

データ研磨とデータリテラシ

データ研磨スキルそのものだけではなく，データリテラシ（データを扱う上で必要な最低限の知識）の欠如も分析結果へ影響を及ぼす．データ分析の前処理において「使用するデータの特性把握」は軽視

されやすい事項である．特に e-Stat 等で公表されている公的統計については，調査の目的や調査対象，利用上の注意まで細かに情報が開示されているにも関わらず，誤ったデータの使い方をするケースが散見される．たとえば経済センサスは 2009 年に開始されたが，その前身は「事業所・企業統計調査」であった．経済センサスでは登記簿情報も活用しているため，事業所・企業統計調査と比べると調査対象が拡大している．総務省統計局の Web サイトでも「国においては統計表の時系列比較を行っておりません．その点を十分にご留意願います．」と明記されている[☆2]．よって，2006 年の事業所・企業統計調査の事業所数と 2009 年の経済センサス基礎調査の事業所数を単純に比較し，「事業所数は増加傾向にある」と結論づけることはできない．このように，データを加工するスキルだけでなく，大前提として分析に用いるデータがどのような仕様であるのかも理解する必要がある．データ研磨やデータ分析を正しく行ったとしても，そもそも比較できない対象を比較して論じてしまえばその分析結果は誤りとなる．仕様の把握だけにとどまらないが，データリテラシは，データサイエンティストやエンジニアだけでなくデータにかかわるすべての人が修得すべき共通言語といえるのではないか．

データ研磨スキルの教育

データ研磨はデータ分析による価値創造の土台であるが，我が国では教育プログラムが確立されていない．統計教育においては分析手法の習得が主となり，データ研磨のノウハウの体系化はあまり重視されてこなかった．データ分析を目的としたプログラミングについては個人の学習による習得が主で，データ分析の現場でのスキルレベルは個人によるばらつきが大きい．教育によるデータエンジニア人材の養成は急務な状況にある．帝国データバンクはビッグデータの加工を行うデータエンジニアの育成によるデータサイエンス分野の発展を目的とし，2017 年 11 月に滋賀大学と連携協力協定を締結した．2015 年 4 月に公開された経済産業省と内閣官房が提供している地域経済分析システム（RESAS）において，公開に先んじて 2014 年以降，帝国データバンクは官民さまざまなデータの研磨を行ってきた．その中で蓄積したノウハウと，滋賀大学データサイエンス学部の教育プラットフォームを組み合わせることで，従来体系化されていなかったデータ研磨の体系化を目指している．

体系化の取り組みの 1 つとして，2018 年 10 月・11 月には，滋賀大学データサイエンス学部の 2 回生 20 名を対象に「データエンジニアリング人材養成演習」として 90 分× 15 コマの集中講義を週に 1 日，計 4 日間で実施した．講師が学生の作成したプログラムを個別にチェックしフィードバックするという体制であるため，講義の受講者数には制限を設けた．教育プログラムの目的設定としては以下の 3 点である．

- 効率的なデータ加工技術の習得
- 分析におけるデータ研磨の重要性の理解
- 作成データに対する検査・報告ができる

これらを講義の狙いとし，「座学によるスキル習得」と「ツーマンセルによるデータ研磨の実践」を合わせた講義形式で実施した．具体的には，総務省「地方財政状況調査関係資料」や国土交通省「不動産取引価格情報」のデータを題材に，プログラミング言語の R を用いて自らデータを取得，研磨，簡易な分析までの一連の流れを 2 人 1 組のペアで取り組むという形式をとった（図 -4，5）．ペアの 2 人の間で作成したデータが完全に一致するまでデータ研磨を行うことでヒューマンエラーを最小限に抑え，研磨したデータの精度を担保するためのツーマンセルである．なお，データ研磨を行う上で，最低限身につけておけばおおむねのデータを加工できる

☆2 https://www.stat.go.jp/data/e-census/2009/kakuho/riyou.html#hikaku

という使用頻度の高いスキルを 20 個ピックアップし，初級スキルと定義づけた．たとえばデータの入力，縦結合，横結合，条件分岐，グループ集計，出力などが初級スキルの例である．また，反復処理やマクロの活用などによる効率的な処理の実現のために必要なスキルを中級スキルと定義づけた．

地方財政のデータでは初級スキルを活用し，税収データと人口データの統合，市区町村統廃合情報の反映，最新市区町村単位でのグループ集計，一人当たり地方税の算出，データ形式を整理して出力，といった工程をプログラミングで行った．不動産取引のデータでは地方財政データの研磨と研磨工程自体は大きく変わらないものの，中級スキルを活用して 8 年分の生データを一括処理することを一番の目的としている．

集中講義形式で 2 人 1 組のツーマンセル形式でデータ研磨を行い，個別学生の作成したプログラムとデータに対してチェックを行い次週の講義でフィードバックを行ったことにより，学生からはプログラミングによるデータ研磨のスキルが着実に身についたという声が多く挙がった．実際に自身で研究やビジネスの領域でデータを研磨する際には加工済みのデータが用意されているケースはほとんどない．正しく前処理が施されたデータを作成できるかどうかは自身のデータ研磨スキルにかかっている．また，少人数制で実施したことにより個別に学生をフォローしやすくスキルアップにつながっていたと感じる反面，オーダーメイドの要素も含むことから座学の講義と比べて多くのリソースが必要であった．そして，この講義は初の取り組みで，毎週講義を実施したが，それは学生にとっても講師にとっても少々負担が大きかったように感じられる．各講義の最後には，次週までの課題として指定したデータをペアで研磨を行うことを課したが，ペアでスケジュールを合わせて作業することの難しさ，コミュニケーションを取りながら進めることの難しさも課題として挙がった．また，学生が課題を実施する期間を 5 日間，講師が提出物をチェックして講義資料に反映する期間が 1 日というスケジュールであったため，講師側にも少なからず負担があった．それを踏まえると，隔週で実施するなどある程度の期間を確保することが望ましいように思う．教育の方法についてはまだ模索中ではあるが，データ分析におけるデータ研磨の重要性と体系化されたデータ研磨スキルの教育が，データエンジニアおよびデータサイエンティストの育成につながり，データ分析を活用した社会の発展へ貢献できるものと考える．

(2019 年 2 月 1 日受付)

■菊川康彬　yasuaki.kikukawa@mail.tdb.co.jp

2010 年慶應義塾大学経済学部卒業．2012 年修士（経済学）．同年より（株）帝国データバンクに勤務．総合研究所にて企業間取引データの分析に従事．その他，内閣府経済社会総合研究所研究協力員，滋賀大学データサイエンス学部非常勤講師を兼務．

■図 -4　集中講義の様子（ペアワーク）

■図 -5　集中講義のスケジュール