「株式会社技術情報協会　月刊　研究開発リーダー　2023年8月号」掲載記事
URL:https://www.gijutu.co.jp/index.htm

◆連　載◆

《連載 全 3 回》

第 3 回　R & D 部門におけるデータ共有，利活用のためのデータの記録，蓄積，分析方法

上島 豊　　（株）キャトルアイ・サイエンス　代表取締役

《PROFILE》

略歴：

1992 年 3 月　大阪大学工学部　原子力工学科　卒業
1997 年 3 月　大阪大学大学院工学研究科　電磁エネルギー工学専攻　博士課程修了
1997 年 4 月　日本原子力研究所　博士研究員
2000 年 4 月　日本原子力研究所　研究職員
2006 年 3 月　日本原子力研究開発機構（旧日本原子力研究所）　退職
2006 年 4 月　キャトルアイ・サイエンス設立　代表取締役　就任

主な参加国家プロジェクト：

文部科学省 e-Japan プロジェクト　「ITBL プロジェクト」,「バイオグリッドプロジェクト」
総務省 JGN プロジェクト　「JGN を使った遠隔分散環境構築」
文部科学省リーディングプロジェクト　「生体細胞機能シミュレーション」

主な受賞歴：

1999 年 6 月　日本原子力研究所　有功賞
「高並列計算機を用いたギガ粒子シミュレーションコードの開発」
2003 年 4 月　第 7 回サイエンス展示・実験ショーアイデアコンテスト文部科学大臣賞
「光速の世界へご招待」
2004 年 12 月　第 1 回理研ベンチマークコンテスト　無差別部門　優勝

主な著作：

培風館「PSE book ―シミュレーション科学における問題解決のための環境（基礎編）」ISBN：456301558X
培風館「PSE book ―シミュレーション科学における問題解決のための環境（応用編）」ISBN：4563015598
培風館『ペタフロップス コンピューティング』ISBN978-4-563-01571-8
臨川書店『視覚とマンガ表現』ISBN978-4-653-04012-5

7　項目名，単位が重要と言うが，今まであまり気にしなくとも分析できているが・・・

前章では，「記録するデータの項目名には，分析で使うべき単位をつけるべき」と題し，記録するデータの項目名や単位は，どのような分析をするかを具体的に考えて，決めるべきであるということを説明した。本章では，「項目名，単位が重要と言うが，今まであまり気にしなくとも分析できているが・・・」と題し，分析のことをあまり気にしないで項目名や単位をつけていても，今まで大きな問題は起こっていないのではないか？本当に重要なのか？と思っている研究者に，「なぜ，そういう問題が起こっていないのか？本当にそのままで問題はないのか？」に関して，論じる。

「分析のことをあまり気にしないで項目名や単位をつけていても，大きな問題は起こっていない」という感覚は，間違っていないということを最初に伝えておく。しかし，項目名や単位に関しては，そこまで細かく考えて決める必要はないと判断を下すのは早計である。皆がなぜ，「分析のことをあまり気にしないで項目名や単位をつけていても，大きな問題は起こっていない」と感じ，実際，問題が起こっていないのかに関して，以下で種明かしをする。

データ分析をするためには，X-Y プロットを描くはずであり，その時，X 軸以外の実験パラメータが一定の実験結果をプロットしていくはずである。前章では，「項目名や単位を適切に設定していないと，X 軸以外の実験パラメータが一定の実験結果を抽出することが難しい」ことを説明した。実際，Excel にたくさんの実験結果があるデータ表の中から X 軸以外の実験パラメータが一定の実験結果を抽出するのは，項目名や単位を適切に設定していないと相当面倒なことは紛れもない事実である。しかし，皆は，この「相当面倒なこと」は，あまり記憶にないはずである。

そもそも「Excel にたくさんの実験結果があるデータ表」が存在していないということもその理由の一つだが，一番大きな理由は，皆が○○特性値の△△依存性を確認するために実験計画（どの実験パラメータを固定し，どの実験パラメータを変動させるか）を立て，実験をしていることである。「○○特性値の△△依存性を確認するために」ということは，X-Y プロットの X 軸が△△で，

Y 軸が○○特性値ということになる。つまり，実験を行おうとする計画段階で，どのような X-Y プロットを描くかは決められていて，そのプロットが描きやすい実験計画が立てられているのである。当然，そのような実験計画では，X 軸以外の実験パラメータが一定の実験になるに決まっているし，記録するときも△△依存性を表現しやすい項目名，単位になるはずである。

例えば，エチレン，プロピレン，ブタン，触媒を混ぜて合成し，その合成物の粘度を測定する実験で，エチレン，プロピレン，ブタンが 1：3：6 の重量比率の時の粘度の触媒依存性を調べるために実験を行うとする。エチレン，プロピレン，ブタンをそれぞれ 100g，300g，600g として，触媒を 0.1g，0.2g，0.5g，1.0g で実験をし，粘度を測定する計画を立てたとする。この時，実験データは，エチレン重量（g），プロピレン重量（g），ブタン重量（g），触媒重量（g），粘度（Pa・s）で記録し，X 軸に触媒重量（g），Y 軸に粘度（Pa・s）として，実験結果をプロットすれば，粘度の触媒依存性はわかるはずである。そもそもこの時点では，エチレン，プロピレン，ブタンは変動させないつもりなので，この実験に関しては，どんな単位でも問題はないのである。

しかし，過去実験でエチレン，プロピレン，ブタンがそれぞれ 10g，30g，60g で，触媒が 0.03g，0.07g，0.5g の結果があった場合に，この項目名，単位ではその結果は同じグラフにプロットできない。本来は，「エチレン，プロピレン，ブタンが 1：3：6 の重量比率の時」というのは，エチレン重量（g），プロピレン重量（g），ブタン重量（g）では，それはそのまま正確には表現できないはずであり，本来は，以下のような濃度単位を定義して項目化すべきである。

エチレン濃度（wt%）
＝エチレン重量（g）/{エチレン重量（g）＋プロピレン重量（g）＋ブタン重量（g）} プロピレン濃度（wt%）
＝プロピレン重量（g）/{エチレン重量（g）＋プロピレン重量(g)＋ブタン重量(g)} ブタン濃度(wt%)
＝ブタン重量（g）/{エチレン重量（g）＋プロピレン重量（g）＋ブタン重量（g）}

また，エチレン，プロピレン，ブタンが重量濃度単位なので，触媒も以下のように重量濃度単位で定義すべきである。

触媒濃度（wt%）
＝触媒重量（g）/{エチレン重量（g）＋プロピレン重量（g）＋ブタン重量（g）}

つまり，今行おうとしている実験だけで問題ない項目名でなく，そもそもどんな依存性を確認しようとしているのかに立ち返り，一般的な実験結果も同じグラフにプロット可能な項目名，単位を考えないといけないということである。実は，これが簡単なようで，研究者の皆ができていないことなのである。

「○○特性値の△△依存性を確認するためにということを決めた段階で項目名や単位を決めればいいし，実際，今まで無意識とはいえ，それができているのであれば，問題ないのでは？」と思われるかもしれない。今まで通りの研究スタイルでいいのであれば，答えは Yes である。しかし，自分が過去行った実験や他研究者が行った実験の結果を使って，「○○特性値の△△依存性を確認する」をしたいのであれば，No である。逆に言うと，皆が「自分の過去データや他研究者のデータは扱いが難しいから，再実験をしないといけなくなる」という問題の原因は，「○○特性値の△△依存性を確認する」と思った，その場その場で，十分な一般化をせず項目名や単位を決めているからである。その結果，項目名や単位が揃わなくなったり，揃っても扱いにくい単位（例えば，g）になっていたりして，自分の過去データや他研究者のデータが扱い難くなるのである。したがって，自分の過去データや他研究者のデータを本気で使おうと思わないのであれば，今まで通りの項目名や単位の決め方でいい。しかしそうでないなら，項目名や単位の決め方を改める必要があるのである。

8 教科書などであまり触れられていない多変量データ分析の重要な注意点

前章では，「項目名，単位が重要と言うが，今まであまり気にしなくとも分析できているが・・・」と題し，分析のことをあまり気にしないで項目名や単位をつけていても，大きな問題は起こっていないと思っている研究者に，「なぜ，そういう問題が起こっていないのか？本当にそのままで問題はないのか？」について説明した。本章では，「教科書などであまり触れられていない多変量データ分析の重要な注意点」と題し，多変量データ分析において，非常に重要にもかかわらず，データ分析の教科書

にはあまり触れられていない注意点に関して，説明する。

世の中では，MI，AI が叫ばれているが，MI，AI はアイデアを提供してくれるという点では心強いが，「それを良いと思った理由」は提示してくれないので，データ分析，理解という観点では無力である。実際，MI，AI を活用するにしても，MI，AI が提示した案について，実データを分析し，その背景理由を探ることを辞めてしまってはいけない。そもそも，それを辞めるなら研究者はいらないということになる。背景理由を探るための最も初歩的なデータ分析方法が義務教育時代から親しんだ X-Y プロットである。たかが，X-Y プロットであるが，されど X-Y プロットである。結局，様々な学術論文でも賑やかしの 3D 可視化や画像などがあったとしても，論文で一番重要な部分は X-Y プロットになっているのは，偶然ではない。

X-Y プロットは，「Y 軸項目値の X 軸項目値依存性」を表すもので，データ分析，理解というものの基本的な部分は，まさしくこれを使う必要がある。一般的に X 軸項目に実験パラメータを割り付け，Y 軸項目に実験結果項目を割り当てる。実は，注目している実験パラメータを X 軸項目に指定し，依存性を確認したい実験結果項目を Y 軸項目に指定しただけでは駄目なのである。実際のデータ分析では，いくつかの注意が必要である。まず，どんな実験でも実験パラメータが一つということはあり得ないはずで，注目している実験パラメータ以外にも実験パラメータがたくさんあるはずである。その実験パラメータを全く無視して，X-Y プロットを描いてはいけない。実際，そのようなことをすると「Y 軸項目値の X 軸項目値依存性」を見ているようで，表には出てきていない X 軸項目以外の実験パラメータ依存性を X 軸項目値依存性と誤認してしまう恐れがある。

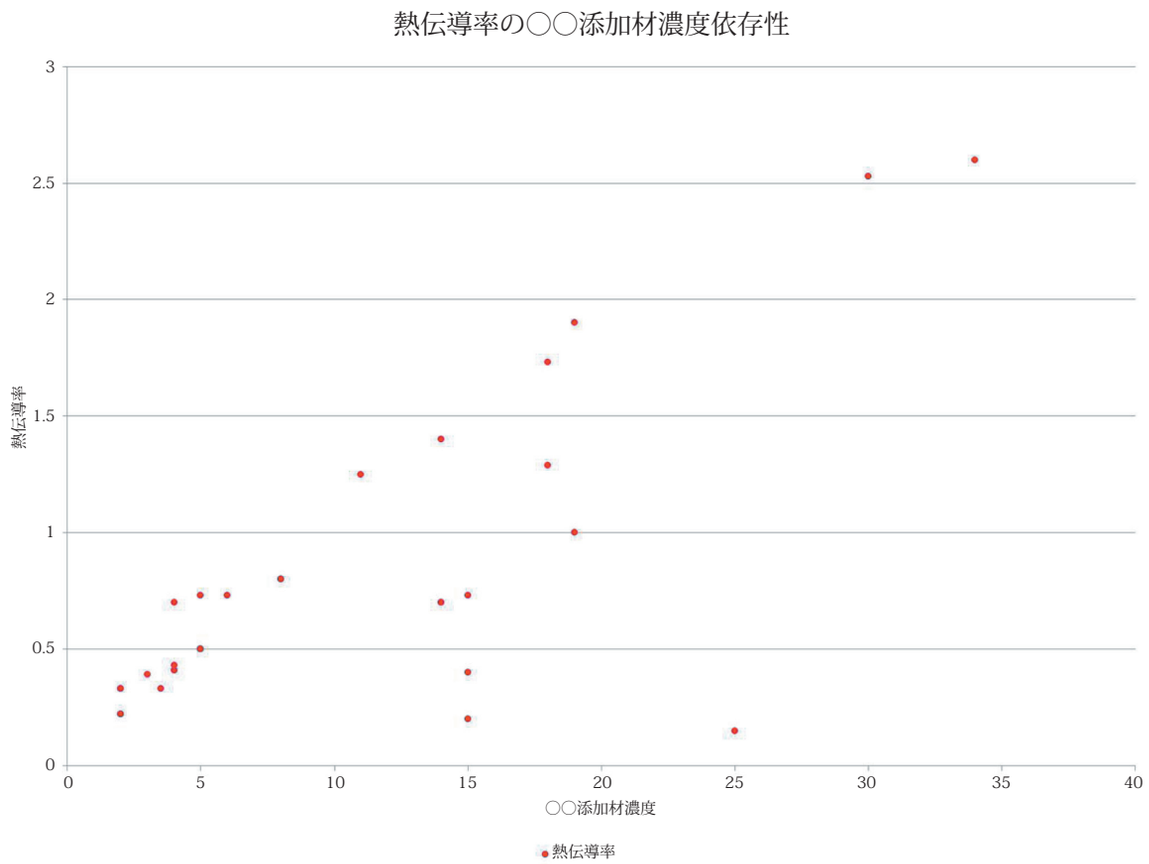

図 1　X 軸以外の実験パラメータが一定であるデータ毎にデータ分類されていない X-Y プロット

図 1 の X-Y プロットは，様々な実験結果に対して，注目している○○添加剤濃度を X 軸項目に指定し，依存性を確認したい熱伝導率を Y 軸項目に指定したもので，添加剤濃度を上げていけば，添加剤濃度 1％あたり，熱伝導度は 0.07 ～ 0.08 ずつ比例して上がっていく傾向にあることが読み取れてしまう。

実は，図 1 の X-Y プロットは問題なのである。「注目している実験パラメータを X 軸項目に指定し，依存性を確認したい実験結果項目を Y 軸項目に指定」する前に X 軸項目以外の実験パラメータが一定であるデータ毎にデータを分類しておく必要があるのである。多工程の実験では，実験パラメータは数百から千程度になるので，「X 軸項目以外の実験パラメータが一定であるデータ毎にデータを分類しておく」作業は，結構大変な作業になる。しかし，この作業を省略して，「Y 軸項目値の X 軸項目値依存性」をグラフ化してはいけない。検討したい X 軸項目以外の実験パラメータが確定している場合は，「X 軸項目以外の実験パラメータが一定であるデータだけにデータを絞り込む」ことは，ひたすら実験パラメータの数だけ Excel の filter をかけていけばいいだけなので，大変な作業にはなるができてしまう。しかし，そもそもその一定にしたい値がよくわかっていない場合は，何らかの方法で「X 軸項目以外の実験パラメータが一定であるデータ」毎に分類しておき，分析をしたい実験パラメータが一定であるデータ塊を選ぶ必要がある。

Excel で数百から数千の filter をかけることも大変だが，「X 軸項目以外の実験パラメータが一定であるデータ」毎に分類するのは，手動処理では永遠に終わらないほどの作業量になってしまう。例えば，Excel で 1000 実験パラメータ項目を「実験パラメータが一定であるデータ毎に分類」しようとすると，1 個目のパラメータ項目の値を filter で 1 個指定し，次のパラメータ項目も値を filter で 1 個指定しという作業を 1000 項目すべてに行い，やっと実験パラメータが一定であるデータが 1 塊だけ分類できる。その後，最初のパラメータ項目の値を別の項目値に filter を変え，同じことを全項目の全項目値のパターンを抜けなく filter をかけていく必要がある。もし，項目値が 2 個ずつだとして，filter を 1 秒で行なえたとしても，全パターンの 2 の 1000 乗パターンを確認するには，2^{1000} 秒＝ 2 の後に 0 が 300 個ぐらい付いた数（秒）＝宇宙の寿命よりはるかに長くかかってしまい，現実的ではない。実際には，何列か filter をかけると結果が 0 件（0 行）になることがほとんどなので，ここまで時間はかからないが，たぶんやりきる人はいないはずである。したがって，多工程の実験を行う R & D 部門では，こういうことが簡単に実施できるツールを整備しておくことが必須ということになる。

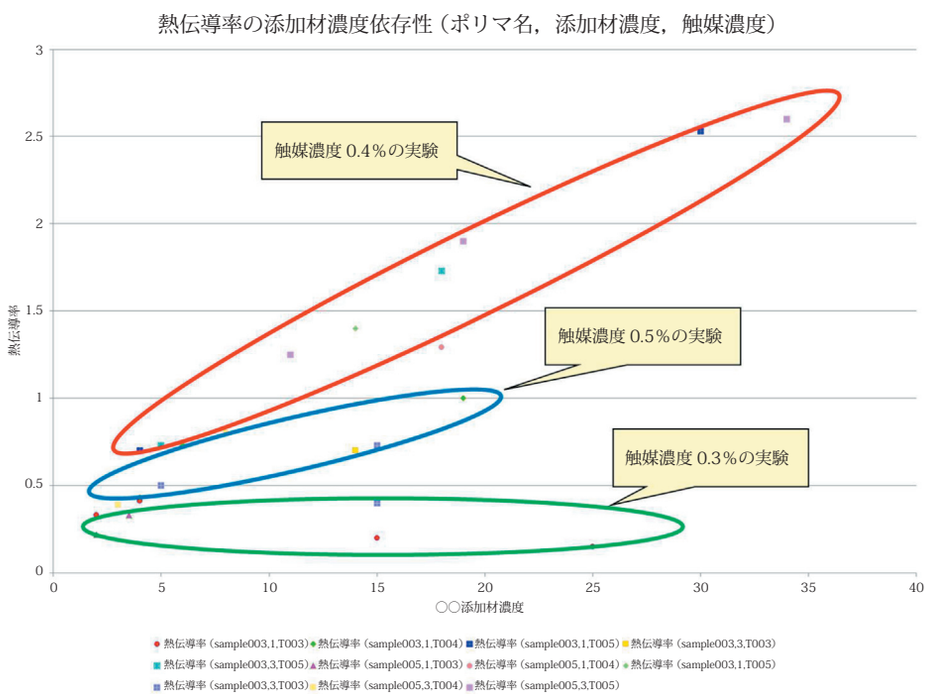

図 2　X 軸以外の実験パラメータが一定であるデータ毎にデータ分類された X-Y プロット

先ほどのデータは，X 軸項目以外の実験パラメータ値で分類をしないで，X-Y プロットを描いたが，図 2 では追加しているポリマー種類や触媒濃度などの X 軸以外の実験パラメータの値が同じデータ毎にデータを分類（マーカー種を分け）し，データをプロットしたものである。そうすると「添加剤濃度 1％あたり，熱伝導度は 0.07 ～ 0.08 ずつ比例して上がっていく傾向がある」のは，触媒濃度を 0.4％にした時だけで，0.3％の場合は，そもそも添加剤濃度依存性がないことがわかる。X 軸以外の実験パラメータの値が同じデータ毎にデータを分類しないと，結論をミスリードしてしまいかねないのである。

実は，データ蓄積，共有を行うまでは，この作業は発生しないか，したとしても大きな負担ではない。データ蓄積，共有を行うまでは，自分が依存性を調べたいと思う実験パラメータ数種だけを変動させて，それ以外の実験パラメータは固定した実験を行うはずであり，また，実験直後にデータ分析を行うはずである。その場合，ほとんどの実験パラメータは固定されており，また，自身も何を固定したという記憶が残っているため，ここで議論している「X 軸の項目以外の実験パラメータが一定であるデータ」を分類することは簡単である。X 軸の項目以外の実験パラメータを一定にして実験をするので，そもそも分類する必要がないことがほとんどである。つまり，こんなグラフを描きたいと思って，それを実現するための実験計画を立てているから，意図したグラフが簡単に描けるのである。

しかし，データ蓄積，共有をし，遥か過去の自分のデータや他研究者のデータを含めて分析をする場合，「どの実験パラメータが固定されているのか？」は，実データを確認するまで分からなく，また，殆どの場合は，実験パラメータの大多数が変動してしまっているはずである。実は，データ蓄積，共有をしない場合は，「実験前にグラフが描きやすいように実験パラメータを固定しているか」，少し前の実験であれば自身の記憶を活用することで「どの実験パラメータが固定されているのか」が分かるので，比較的楽にデータ分析ができるだけなのである。データ蓄積，共有を始めると自身の記憶を使わずにデータ分析をすることが強いられ，その分析の大きな変化を研究者が認識できておらず，当惑してしまうのである。データ蓄積，共有をする前に，自身の記憶を活用せず，実データをしっかり確認しながらデータ分析をする習慣をつけていくことが重要なのである。

データ分析を行うときに，X 軸に実験パラメータでなく，特性値を設定することもあるはずである。例えば，試作物の粘度と熱伝導度の関係性を知りたい場合は，X 軸項目に粘度を設定し，Y 軸項目に熱伝導度を設定して，X-Y プロットを描くはずである。ただ，この時にも幾つか気を付けなければならないことがある。特性値は実験パラメータのように項目値が実験パラメータに対して一対一の関係のある項目でない為，「粘度が○○の時に熱伝導度が××」というプロットがあっても，あらゆる実験でそれが成り立つかと言えば違う。実験パラメータが違っても「粘度が○○」になることは，十二分にあり得て，その時，「熱伝導度が××」にならないことも十二分にあり得るのである。だからと言って，「粘度が○○」になるあらゆる実験パラメータを実験で確かめ，「粘度が○○」になる実験パラメータのすべてで，「熱伝導度が××」になることを調べるのは，現実的でない。もちろん，こういうことは，皆は承知で，いくつかの異なる実験パラメータの結果で，この手のグラフを描いているはずだが，どんな実験パラメータでもその傾向が同じかどうかはしっかりと確かめられていないことが多いはずである。つまり，たまたま，偏ったパラメータの実験での限定的な依存性（偽依存性）を一般的な依存性と勘違いしてしまう可能性が高いのである。したがって，この手のグラフを描く時には，事前に実験パラメータと当該 X 軸，Y 軸に設定する予定の特性値の相関を事前に分析しておき，どのパラメータ領域での粘度と熱伝導度の関係性を評価しようとしているかを明確化しておく必要があるのである。

参考文献

1）上島豊，他
ケムインフォマティックスにおけるデータ収集の最適化と解析手法，技術情報協会，(2023)，p39-74