◆連　載◆

《隔月連載 全 5 回》　　最終回

R & D 部門における機械学習・AI・生成 AI 活用へのデータ共有の重要性

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 ー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 ー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

11　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，「そもそもデータ蓄積，共有は何のために行うのか？」と題し，データ蓄積，共有は AI や機械学習などのために行うのではなく，データ探査，データ処理に網羅性，均質性，再現性を付与することで，研究の品質，客観性を担保するために必要であるということを説明した。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。前章でも少し触れたが，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。

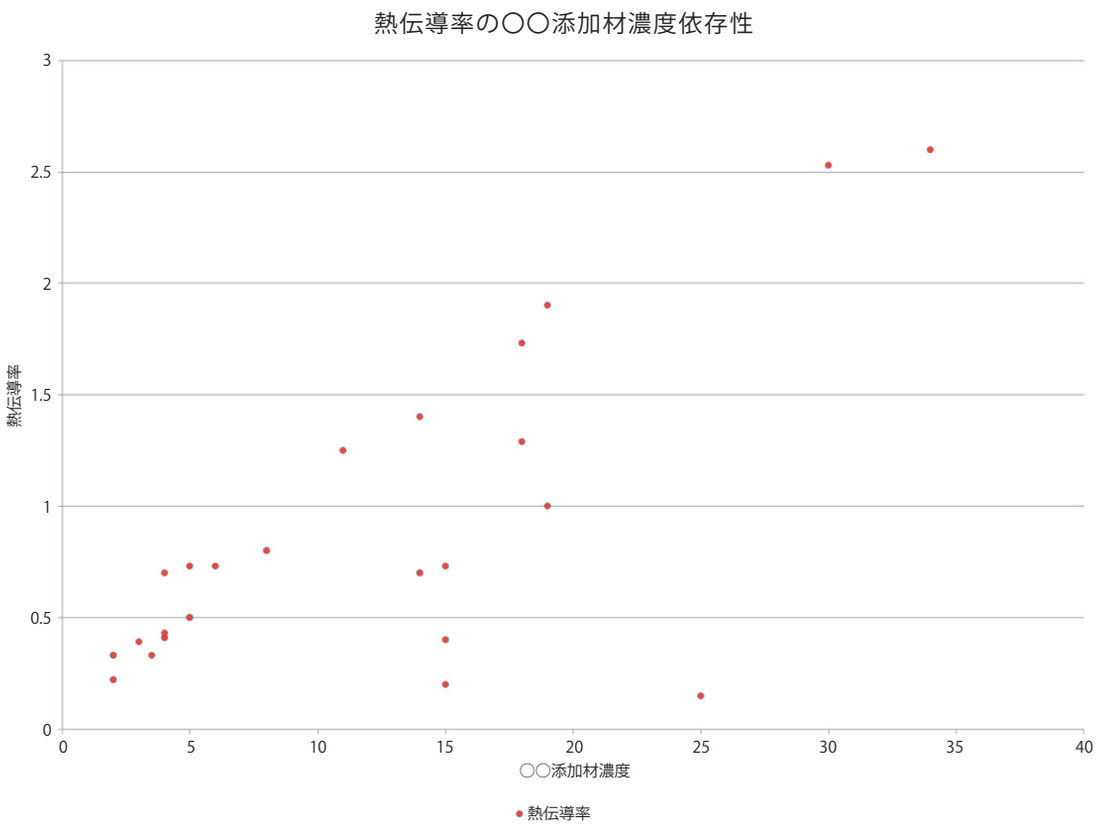

図 27　X 軸以外の実験パラメータが一定であるデータ毎にデータ分類されていない X-Y プロット

上記 X-Y プロットは，様々な実験結果に対して，注目している〇〇添加剤濃度を 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 27 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 個指定してという作業を 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*1 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。

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 28 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，

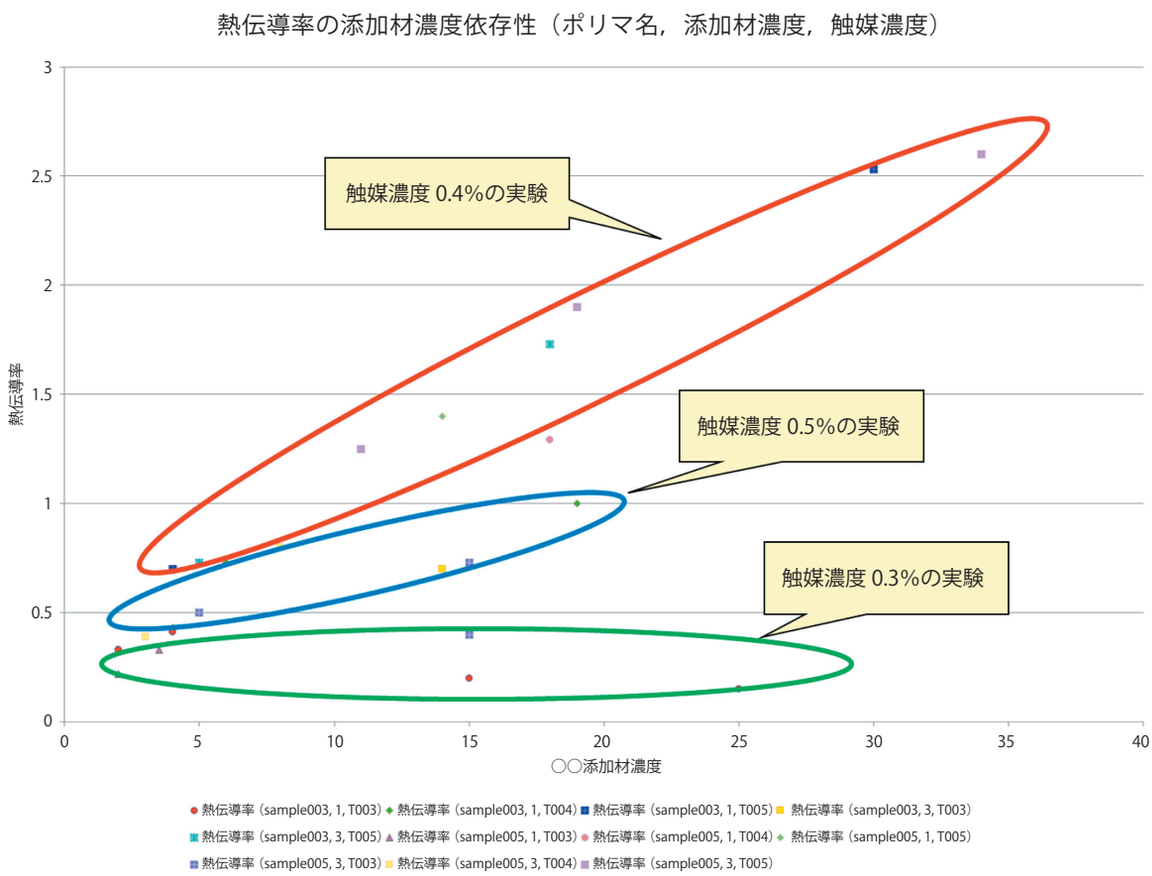

図 28　X 軸以外の実験パラメータが一定であるデータ毎にデータ分類された X-Y プロット

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。実際，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から，そのグラフが簡単に描けるような実験をしているということになる。

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。

こういう話をするときに，「折角データベースがあるのであれば，自分が確認したい実験パラメータ（それこそ，X 軸以外の殆どの実験パラメータ）を検索条件に指定し，その結果を X-Y プロットするのであれば，データ分類（検索条件で既に分類されている）もいらないので，簡単にグラフが描けるのではないか？」という人もいる。そうすれば確かにデータ分類はいらなくなるが，本当にそれで，データ蓄積，共有をしたメリットが出るだろうか？実は検索条件でデータ分類が不要なぐらいに絞り込みをすると自分の直近のデータしか抽出されないことになるのである。X 軸以外の実験パラメータすべてが同じ値になる実験なんて，過去の自分でもしていないはずである。実は，データ蓄積，共有をする本懐は，自分が実施していない，意識していない実験パラメータの実験結果と自分が意識して実施した実験結果を並べ，そこから広いパラメータ空間での関連性を発見することなのである。

自分は添加剤濃度依存性を確認したいと思っていて実験をしていた時に，過去の経験から触媒濃度を 0.3 で十分と判断し，触媒濃度を固定して実験をしていたとする。自分が確認したい実験パラメータ（それこそ，X 軸以外の殆どの実験パラメータ：触媒濃度は当然 0.3）を検索条件に指定し，その結果を X-Y プロットすると，熱伝導率は添加剤濃度に依存しないという結論になってしまう。データ蓄積，共有をするときには，検索条件はあまり厳しく設定せず，自分が意識できていない実験パラメータでの実験結果も含めて傾向を把握しようという態度が重要なのである。これは，今までのデータ分析の傾向の見方である局所依存性とは反対の大域構造の把握であり，頭の考え方もコペルニクス的転回が必要ということになる。

さて，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。

12 実験パラメータ項目は，どのような独立変数系にするかが重要

前章では，「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」に関して，説明を行った。本章では，「実験パラメータ項目は，どのような独立変数系にするかが重要」と題し，実験パラメータ項目に関しては，単なる単独の項目だけを見て，項目名を決めるのではなく，全体を俯瞰し，データ分析のためにはどのような独立変数系が適切かを考える必要があるということを論じる。

具体的な実験パラメータ項目を想像しやすいように，図 29 のような仮想的な実験を考える。ここでは，液体としてモノマー a，モノマー b，触媒 c，気体としてモノマー c，モノマー d，水素 h を温度，圧力が一定の容器に密閉し，容器内でバブリングすることで気液混合をし，合成反応を行う実験を仮定している。

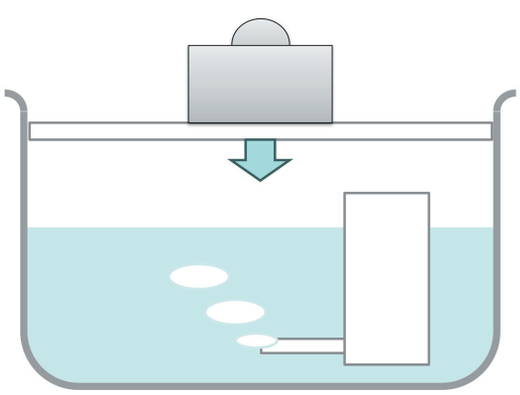

図 29　仮想的な実験環境のイメージ

この時の実験パラメータは，実験の計量，設定を念頭に置き，前章のルールを加味すると，以下のような項目名が自然である。ちなみに，この実験パラメータの独立変数の数は 8 個になる。

m_a_ 液体モノマー a 使用量（g）
m_b_ 液体モノマー b 使用量（g）
m_c_ 気体モノマー c 使用量（g）
m_d_ 気体モノマー d 使用量（g）
m_h_ 水素ガス使用量（g）
m_s_ 液体触媒 s 使用量（g）
P_ 容器内圧力（MPa）
T_ 容器内温度（℃）

ここで，実験結果として得られる合成物分子量が，容器内温度に対してどのような依存性を持つのかを蓄積データから調べたいとする。X 軸に T_ 容器内温度（℃）を設定し，Y 軸に合成物分子量を設定した X-Y プロットを描けば，依存性は分かるはずである。

実は，この言明は半分は合っているが，半分は間違っている。合成物分子量の容器内温度依存性を明らかにしようとし，今まさに実験をする場合は，たぶん，容器内温度以外の実験パラメータ（各種使用量と圧力）を固定して，容器内温度だけを変えた実験をいくつか行うはずである。したがって，上記 X-Y プロットを描けば，容器内温度依存性は分かるはずである。しかし，過去に様々な実験パラメータで実施し，蓄積されているデータでは，容器内温度以外の実験パラメータもバラバラなはずであり，そのまま X-Y プロットを描いても，容器内温度依存性は分からないはずである。・・・と言うか，グラフを描いてしまうと，依存性らしきものが見えることもあり，容器内温度依存性を読み取ってしまうことの方が危険である。依存性を見て取るには，図 30 のように X 軸以外の実験パラメータの値が同じデータごとに分類し，それぞれのデータごとに X-Y プロットを描く必要があるのである。

さて，図 30 を見て，何かが分かるだろうか？実は，蓄積データを使ったデータ分析で最初に躓くのは，ここなのである。せっかく項目名を揃えてデータ蓄積をしても，このように意味の分からない大量のグラフが生成され，そこからどのように検討を進めればいいのか分からなくなる。そして，分からないので，機械学習や AI を勉強して・・・というのがよく聞く流れである。ここで，強調しておくが，X-Y プロットで読み取れないことが機械学習や AI で分かるようになると考えるのは危険で，それであれば分析のための研究者は不要で，実験を精緻に実施できる技術者がいればいいだけということになる。X-Y プロットでの読み取り補助のために機械学習や AI を使うべきなのである。社会科学のようにそもそも対象が相当複雑で，実験パラメータを振った人為的実験ができない分野では，X-Y プロットでの読み取り自体も主観的にならざるを得ないので，機械学習や AI に主観排除としての大きな意味がある。しかし，実験パラメータを任意に設定し，他の環境要因を排除できる実験が可能な工学領域の R & D では，データ分析に客観的意味，解釈が十分付与できるので，X-Y プロットでの読み

図 30　合成物分子量の容器内温度依存性を示す X-Y プロット

取りを省略することはあり得ない。話が脱線してしまったので話を少し戻す。図 30 から傾向が把握できないのは，実は X 軸以外の実験パラメータの項目がデータ分析に適していないからなのである。

図 30 の実験パラメータ項目（＝独立変数）は，原材料 6 個に対して，全て「○○使用量（g）」となっている。当然この実験では，何らかの化学反応が起こるはずなのであるが，そもそもその基礎となる化学反応式は，絶対量に対する式ではない。化学反応を決めているのは，絶対量ではなく比率である。したがって，実験パラメータ項目は，以下のように比率で定義した方がデータ分析は行い易いはずである。もちろん，ナノ粒子反応や界面反応を考慮すると絶対量に依存する化学反応もある。しかし，今，注目しているところが絶対量に依存する化学反応でないのであれば，比率を項目にした方が分析はし易いはずである。変数を変換しているだけなので，比率の方が分析がし易いだけであって，ナノ反応や界面反応を無視した分析になってしまうわけではないことを付け加えておく。

モノマー b 液体部数　　＝ m_b / m_a
触媒 s 液体部数　　＝ m_s / m_a
モノマー d 気体部数　　＝ m_d / m_c
水素ガス h 気体部数　　＝ m_h / m_c
総重量（g）＝ m_total　＝ m_a ＋ m_b ＋ m_s ＋ m_c ＋ m_d ＋ m_h
気体液体重量比＝ r_g_l ＝（m_c ＋ m_d ＋ m_h）/（m_a ＋ m_b ＋ m_s）
P_ 容器内圧力（MPa）
T_ 容器内温度（℃）

実際，m_total 総重量（g）や r_g_l 気体液体重量比が十分大きなとこでは，m_total 総重量（g）や r_g_l 気体液体重量比の合成物分子量に対する依存性はなくなるはずである。m_total 総重量（g）が十分大きなところでは，表面反応やナノ粒子反応は無視でき，r_g_l 気体液体重量比が十分大きなところでは気体は液体に対して飽和濃度に達し，液体中反応は変化しないと考えられるからである。そういう物理傾向を捉えた変数系を実験パラメータ項目にしておくと，データ抽出，分類，分析が容易になる。先ほどの変数系で，X 軸に総重量を設定し，X 軸以外の実験パラメータの値が同じデータごとに分類し，それぞれのデータごとに X-Y プロットを描くと図 31 のようになり，ごく一部の実験を除いて，総重量依存性がないことが確かめられる（はずである）。

図 31　合成物分子量の総重量依存性を示す X-Y プロット

さらに，X 軸を気体液体重量比に変え，X 軸以外の実験パラメータの値が同じデータごとに分類し，それぞれのデータごとに X-Y プロットを描くと図 32 のようになり，一部の実験を除いて，気体液体重量比依存性がないことが確かめられる（はずである）。

ここまでで確かめられた依存性がない部分のデータだけで，X 軸を温度，X 軸以外の実験パラメータ（依存性がない部分なので，総重量と気体液体重量比を除く）の値が同じデータごとに分類し，それぞれのデータごとに X-Y プロットを描くと図 33 のようになる。気体，液体の材料の比率毎の合成物分子量の容器内温度依存性がよく分かる形になっていることが分かると思う。

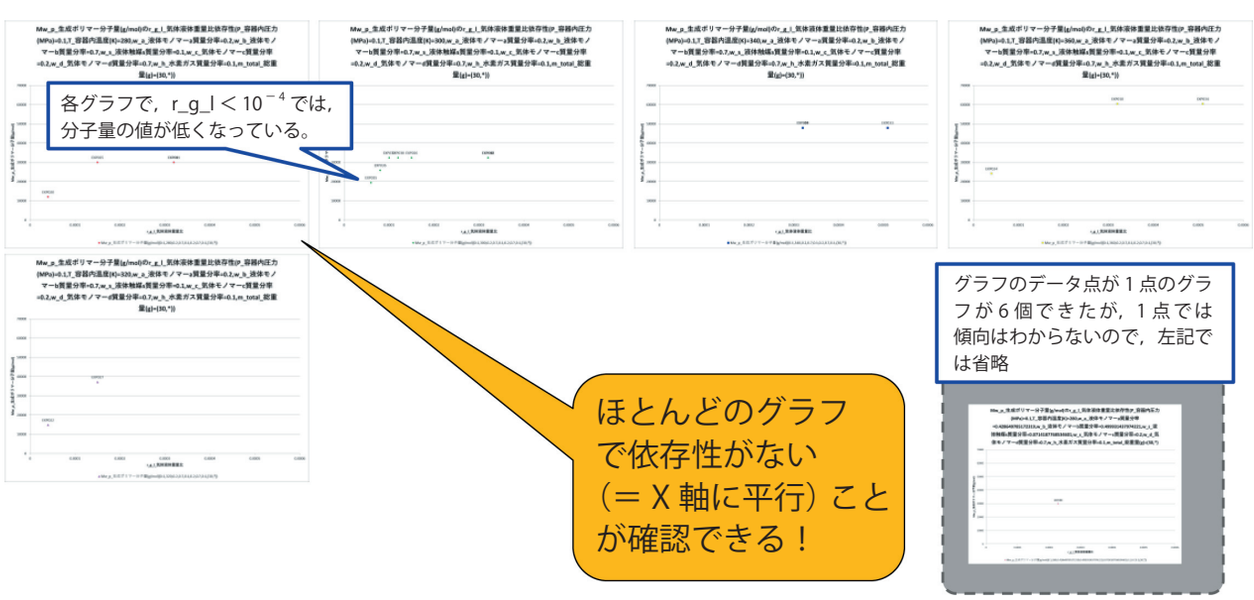

図 32　合成物分子量の気体液体重量比率依存性を示す X-Y プロット

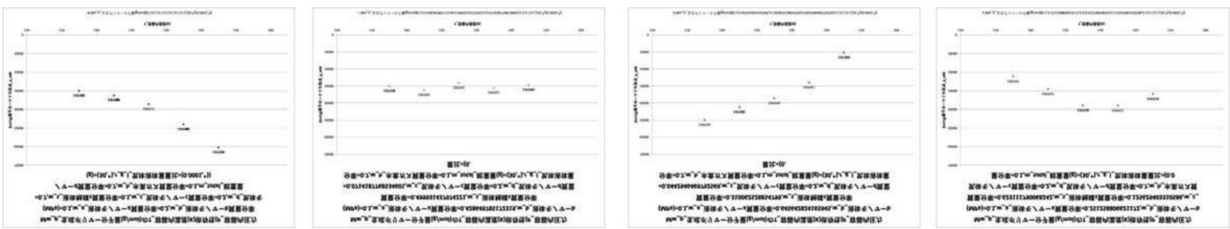

図 33　総重量，気体液体重量比依存性がないデータでの合成物分子量の容器内温度依存性を示す X-Y プロット

「どのような独立変数系にするかをあまり気にしないでデータ分析をしても，大きな問題は起こっていない」という感覚を持っている人はいないだろうか？実は，その感覚自体は間違ってはいない。だからといって，どのような独立変数系にするかに関しては，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，「どのような独立変数系にするかをあまり気にしないでデータ分析をしても，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，以下で理由を説明する。

前章でも触れたが，合成物分子量の容器内温度依存性を明らかにしようとし，今まさに実験をする場合は，「たぶん，容器内温度以外の実験パラメータを固定して，容器内温度だけを変えた実験を行うので，全てのデータで X 軸以外の実験パラメータが固定されている」のである。もちろん，X 軸以外の実験パラメータが 1 パターンだけでなく，数パターンある場合もあるだろうが，それぞれで分析すればいいだけである。ここで，注目すべきなのは，「重量が一定なら重量比は必ず一定になる」が，「重量比が一定だからと言って重量は必ずしも一定ではない」という点である。自分で実験をするときは，重量を一定にして実験をすると重量比も自ずと一定になるので，データ分析の観点では，重量でも重量比でもどちらでも良いのである。しかし，蓄積されたデータを使って分析をするときは，変数系が重量では問題である。蓄積されたデータ全ての重量が一定に揃ったデータではないので，データ分析をするためには，様々な重量パターンから，同じ重量比のデータを分類しなおす必要がある。全て 1g の実験も全て 10g の実験も比が同じなので，「総重量依存以外では分析上は区別したくない」ということである。

したがって，変数系は重量でなく，重量比である必要がある。つまり，データ蓄積，共有をして初めて変数系の問題が噴出するのである。そして，これがデータ蓄積をする前に，変数系の問題が認識されなく，実際に蓄積し出してから実験者が蓄積データを有効に利用できない原因の一つである。

最後に，蓄積データ分析及び蓄積する実験自体に関して，いくつかアドバイスをしておく。蓄積されたデータを分析する場合は，今回分析をしたように通常，依存性がないと思われる実験パラメータに関しても，依存性が認められる領域の実験もあることを前提に分析をすべきである。実は，このような分析を行うことで依存性が認められる領域が見つかるということだけでなく，外れ値から失敗実験（依存性がないと思われるところで依存性がある）も検知でき，今まで何となく外れ点として除去していた実験をしっかりした理由で，除去することができる。

また，新規実験や追加実験を行う場合，総重量，気体液体重量比などの依存性がない領域がどの辺りかをある程度確かめた後は，それら依存性のない（総重量，気体液体重量比等）実験パラメータは固定して実験をすることで，効率的かつ分析精度の高い実験が可能になることも覚えておくとよい。

最後に少しだけ，付け加えておく。実は適切な変数系というのを見出すのは結構骨の折れる仕事であり，以下の変数系を思いつくためには結構深い洞察力が必要である。

モノマー b 液体部数　＝ m_b / m_a
触媒 s 液体部数　＝ m_s / m_a
モノマー d 気体部数　＝ m_d / m_c
水素ガス h 気体部数　＝ m_h / m_c
総重量（g）＝ m_total　＝ m_a ＋ m_b ＋ m_s ＋ m_c ＋ m_d ＋ m_h
気体液体重量比＝ r_g_l　＝（ m_c ＋ m_d ＋ m_h ）/（ m_a ＋ m_b ＋ m_s ）

あまり深く考えない場合は，以下のような変数系にしてしまうのではないだろうか？

モノマー a 液体重量%
= 100 × m_a / (m_a + m_b + m_s + m_c + m_d + m_h)
モノマー b 液体重量%
= 100 × m_b / (m_a + m_b + m_s + m_c + m_d + m_h)
触媒 s 液体重量%
= 100 × m_s / (m_a + m_b + m_s + m_c + m_d + m_h)
モノマー c 気体重量%
= 100 × m_c / (m_a + m_b + m_s + m_c + m_d + m_h)
モノマー d 気体重量%
= 100 × m_d / (m_a + m_b + m_s + m_c + m_d + m_h)
水素ガス h 気体重量%
= 100 × m_h / (m_a + m_b + m_s + m_c + m_d + m_h)

まず，この変数系は変数変換としては間違っている。変数の数は確かに変換前後で同数（6 個）だが，重量%はすべての変数の和は，100 になるので，独立変数ではなくなっている。この変数系では，独立変数として，総重量が必要なのである。

総重量 (g) = m_total
= m_a + m_b + m_s + m_c + m_d + m_h

まあ，この点は気づいている人の方が多いと思うが，問題はこの先である。例えば，合成物分子量のモノマー b 液体重量%依存性は，どのようになるかを調べたい場合はどうすればよいのだろうか？ X 軸にモノマー b 液体重量%を設定し，Y 軸に合成物分子量を設定した X-Y プロットを描けばいいのではないか？もちろん，11 章で説明した X 軸以外の実験パラメータの値で分類をして・・・。本当にそれだけで，大丈夫だろうか？ X 軸以外の実験パラメータは何になるのだろうか？モノマー a 液体重量%, 触媒 s 液体重量%, モノマー c 気体重量%, モノマー d 気体重量%，水素ガス h 気体重量%，総重量 (g) だろうか？

モノマー a 液体重量%，触媒 s 液体重量%，モノマー c 気体重量%，モノマー d 気体重量%，水素ガス h 気体重量%が決まってしまえば，モノマー b 液体重量%の値も決まってしまう。つまり，それではモノマー b 液体重量%依存性なんて確認できないのである。それは，すべての重量%変数の和は 100 になるからそんなことになるだけで，何か一つ変数を削ればいいだけだと思っている人も多いのではないだろうか？例えば，モノマー a 液体重量%を削ってみよう。そうすれば，モノマー b 液体重量%依存性のグラフは確かに描ける。しかし，モノマー b 液体重量%が大きくなるとモノマー a 液体重量%は小さくなる。このグラフは，純粋にモノマー b 液体重量%依存性を表しているのだろうか？

少し頭を冷やして，モノマー b 液体重量%依存性というのは，どういうものを見たいと思っているのかを考えてみる。まず，「モノマー b 液体重量%以外の実験パラメータは固定で依存性を見る必要がある」という言葉を少し深く考察してみよう。「モノマー b 液体重量%以外の実験パラメータが固定」といった時点では，「モノマー a 液体重量%，触媒 s 液体重量%，モノマー c 気体重量%，モノマー d 気体重量%，水素ガス h 気体重量%のそれぞれが固定」ということを明確に意識しているだろうか？これらを固定するとモノマー b 液体重量%も 1 つに値しかとらなくなってしまうので，少し考えるとそうではないことに気づくはずである。問題はここでも，「少し考えない」と気づけないというのが厄介ではある。

こういう場合は，抽象的に物事を考えるとどうあるべきかが見えにくくなってしまうので，少し具体的な例で考えてみると良い。重量%が問題をややこしくしている原因のようなので，まず，重量で考えてみよう。モノマー b 液体重量依存性を確認する一因子実験を考えてみると，モノマー a 液体重量，触媒 s 液体重量，モノマー c 気体重量，モノマー d 気体重量，水素ガス h 気体重量を一定にし，モノマー b 液体重量だけ増減させる実験を行うのではないだろうか？ここで少し抽象度をあげてみる「モノマー a 液体重量，触媒 s 液体重量，モノマー c 気体重量，モノマー d 気体重量，水素ガス h 気体重量を一定にし」は，本当にそれぞれの重量を一定にすることに実験としてそこまでの意味があるだろうか？実際，一定にしたいのは，「モノマー a 液体重量，触媒 s 液体重量，モノマー c 気体重量，モノマー d 気体重量，水素ガス h 気体重量の比率」ではないだろうか？実は，モノマー b 液体重量依存性を確認する一因子実験で一定にしたい，すべきものは，モノマー a 液体重量：触媒 s 液体重量：モノマー c 気体重量：モノマー d 気体重量：水素ガス h 気体重量の連比なのである。連比が一定と

いうのは，触媒 s 液体重量 / モノマー a 液体重量，モノマー c 気体重量 / モノマー a 液体重量，モノマー d 気体重量 / モノマー a 液体重量，水素ガス h 気体重量 / モノマー a 液体重量が一定ということと等価である。そうするとモノマー b 液体重量依存性というのも，モノマー b 液体重量 / モノマー a 液体重量依存性ということになる。つまり，モノマー b 液体重量％依存性を確認したいということは，深く考えると「モノマー b 液体重量 / モノマー a 液体重量依存性を確認したい。その時，モノマー b 液体重量 / モノマー a 液体重量以外の変数：触媒 s 液体重量 / モノマー a 液体重量，モノマー c 気体重量 / モノマー a 液体重量，モノマー d 気体重量 / モノマー a 液体重量，水素ガス h 気体重量 / モノマー a 液体重量及び全重量を一定にして」ということなのである。

比は重量であろうと重量％であろうと同じ値になることや重量％というのは，なんとなく比率という感覚でとらえているのが，混乱のもう一つの原因なのである。同じ比率でもモノマー a 液体重量％，触媒 s 液体重量％，モノマー c 気体重量％，モノマー d 気体重量％，水素ガス h 気体重量％が一定ということと，モノマー a 液体重量％：触媒 s 液体重量％：モノマー c 気体重量％：モノマー d 気体重量％：水素ガス h 気体重量％が一定というのは，意味が違うのだが，相当意識しないとその違いに気づかないのである。

結局何が言いたいかというと，重量％や濃度というのは日常でも良くつかわれる単語で，日常で使う場合には同じ単語を場面，場面で少しずつ違う定義で使っていることがあるのである。実はよく考えるとモノマー b 液体重量％依存性という言葉自体，実際にはその時頭で意図しているものではないのである。頭にあるのは，モノマー b 液体重量相対比率依存性で，モノマー a 液体重量：，モノマー b 液体重量：触媒 s 液体重量：モノマー c 気体重量：モノマー d 気体重量：水素ガス h 気体重量：モノマー b 液体重量の連比で，モノマー b 液体重量の比率だけを変化させた場合を意識しているということなのである。そして，この連比を個々の比率にすると本章の最初の方でデータ分析に使った

モノマー b 液体部数　＝ m_b / m_a
触媒 s 液体部数　＝ m_s / m_a
モノマー d 気体部数　＝ m_d / m_c
水素ガス h 気体部数　＝ m_h / m_c

のようになるのである。

13　おわりに

本連載では，R & D 部門における機械学習・AI・生成 AI 活用の可能性と限界について説明をし，それらを活用するために必要なデータ蓄積，共有に関して，実情と改善方法を示した。また，蓄積データの分析は，従来の自分が行ったデータの分析と大きく異なる点があることを実例ともに紹介し，機械学習・AI を使う場合には，そのような従来型の分析も行っていく必要があることを説いた。読んでいただいてお分かりだと思うが，R & D のデータ共有を行なうということは，何かを買って，解決するという問題ではなく，企業風土及び R & D 従事者の意識変革が必要である。R & D 従事者は，業務内容が極めて専門的で，会社内でも対話する機会が少ない人達だと思う。研究者は，とっつきにくいところはあるが，地頭は良く，非常にまじめなので，これを機に話をしてみれば，仲良くなり，本連載で話したようなことを一緒に議論できるかもしれない。そういう人とのつながりを作ることが，R & D 部門におけるデータ共有の第一歩である。

参考文献

1）川田重夫，田子精男，梅谷征雄，南多善，上島豊，他
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），川田重夫，田子精男，梅谷征雄，南多善　共編，培風館，（2005），p69-82

2）谷啓二，奥田洋司，福井義成，上島豊
ペタフロップスコンピューティング，
矢川元基　監修，培風館，（2007），p183-202

3）牧野圭一，上島豊，視覚とマンガ表現，臨川書店，（2007），p1-5，221-229

4）上島豊，月刊「研究開発リーダー」6 月号，技術情報協会，（2023），p63-68

5）上島豊，月刊「研究開発リーダー」7 月号，技術情報協会，（2023），p86-91

6）上島豊，月刊「研究開発リーダー」8 月号，技術情報協会，（2023），p78-82

7）上島豊，月刊「研究開発リーダー」7 月号，技術情報協会，（2024），p77-84

8）上島豊，月刊「研究開発リーダー」9 月号，技術情報協会，（2024），p73-81

9）上島豊, 月刊「研究開発リーダー」11 月号, 技術情報協会, (2024), p85-96
10）上島豊, 月刊「研究開発リーダー」1 月号, 技術情報協会, (2025), p74-82
11）上島豊, 月刊「研究開発リーダー」3 月号, 技術情報協会, (2025), p68-73
12）上島豊，他
研究開発部門への DX 導入による R & D の効率化，実験の短縮化，技術情報協会，(2022)，p195-221
13）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
14）上島豊，他
実験の自動化・自律化による R & D の効率化と運用方法，技術情報協会，(2023)，p159-199
15）上島豊，他
少ないデータによる AI・機械学習の進め方と精度向上，説明可能な AI 開発，技術情報協会，(2024)，p112-127