◆連　載◆

《隔月連載 全 5 回》

第 4 回　R & D 部門におけるデータ共有，AI 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

11　記録するデータの項目名は，意味が分かれば何でも良いという訳ではない

前章では，「R & D 部門においてデータベースを機能拡張していく場合の注意点」に関して，データベースを機能拡張していく場合に陥りがちな問題とその回避方法の説明を行った。本章では，より具体的なデータ蓄積の課題として，「記録するデータの項目名は，意味が分かれば何でも良いという訳ではない」と題し，記録するデータの項目名をどのように決めるべきかを論じる。

項目名決定において重要な点は，データ分析を前提にし，データ抽出／検索，分析を行いやすい項目名にしておかないといけないということである。データ検索，分析を行いにくい項目名だと，結局，データを探したり，分析したりするのに大きな手間がかかるため，データ共有を大きく阻害するのである。R & D 部門以外の他の業務系の場合は，項目名決定時にデータ分析のことはあまり考えず，どのようにデータを入力，登録するかを考え，項目名を決定する。つまり，データ入力のしやすさを前提にして，項目名を決定するのである。R & D 部門以外の他の業務系の場合は，データ検索，分析が複雑ではないので，データ入力のしやすさを前提にして，項目名を決定しても大きな問題にはならないのである。一方，R & D 部門では，項目が多く，項目の追加や変更も頻繁にあるなかで，複雑なデータ検索，分析が必要となるので，データ抽出／検索，分析を行いやすい項目名にしないとデータ抽出／検索，分析自体が困難になる。以下，R & D 部門でデータを記録していくための項目名を決めるときに注意すべき点を列挙しておく。

11.1　すべてのデータは，項目名－項目値という単純構造に落とし込む

「すべてのデータは，項目名ー項目値という単純構造に落とし込む」は，「データを分析するときには，2 次元の表形式データになっている必要がある」ことからくる条件である。Excel などで，グラフを描いてデータ分析をするときには，必ず項目名－項目値の形になっていることは理解できると思う。しかし，これは結構守られていないのである。例えば，実験ノートや実験データを記録する Excel では，以下のような表になっていることが多いと思う。

乾燥工程				焼結工程	
1 段目		2 段目			
温度	風速	温度	風速	温度	風速

実は，この状態は項目名－項目値の単純構造にはなっていないのである。「純粋な項目名は温度と風速だけで，乾燥工程，焼結工程や 1 段目，2 段目などは，分類的なもので項目名でない」と捉えていてはだめなのである。もし，そう考えてしまうと温度＝ 120 という項目名－項目値を見たときに，それが乾燥工程の温度なのか？焼結工程の温度なのか区別がつかない。項目名は重複がなくユニークな命名でないと，データ分析時にそれが何の値なのかが厳密に分からず，困ってしまうのである。つまり，乾燥工程，焼結工程や 1 段目，2 段目などの分類的な情報を含めて項目名にしなければならないのである。Excel でセル結合を使っているデータ記録は，この段階で失格ということになる。結局，先ほどのデータを項目名－項目値の単純構造にすると，以下のようになる。

つまり，Excel で言うと 1 行目の 1 セルずつに項目名が列挙され，2 列目以降は項目値だけが並び，それ以外の分類情報などは一切無い形が，項目名－項目値の単純構造の項目ということである。「Excel でセル結合を使っていても，ピボットテーブルやマクロを使えば，データ分析はできるのでは？」という方もいらっしゃると思う。もちろん，それは間違っていない。ただし，R & D 部門で取り扱う様々な実験すべてに対応するようにできるかというと非常に悩ましいはずである。ピボットテーブルやマクロは，ある一定の決まりきったことを何度も行う場合は，便利なのだが，そもそもそれを作るのが大変なのである。経理や営業などの業務系で，いつも同じ項目データで同じ処理をするのであれば，ピボットテーブルマクロは持続的運用が可能なのだが，扱う材料や処理プロセスがどんどん変わっていく R & D 部門では，ピボットテーブルやマクロを作成し，メンテナンスすることの負担が大きく，現実的には運用できなくなってしまうのである。実際，使われなくなったマクロや GUI が施された Excel が乱立し，収拾がつかなくなってしまった経験のある人も多いのではないかと思う。R & D 部門のデータ記録は，単純で「それでいいのか？」と思ってしまうが，項目名－項目値の単純構造でなければならないのである。「Simple is best.」である。

1 段乾燥温度	1 段乾燥風速	2 段乾燥温度	2 段乾燥風速	焼結温度	焼結風速

実験 ID	原材料名 1	原材料濃度 1	原材料名 2	原材料濃度 2	引張強度
EXP1	エチレン	80	プロピレン	20	10
EXP2	ブタン	75	エチレン	25	12
EXP3	プロピレン	60	ブタン	40	8

11.2　項目名－項目値の項目間に論理的関係があってはいけない

「項目名－項目値の項目間に論理的関係があってはいけない」というタイトルにしているが，これでは何を言っているのかピンとこない人も多いと思う。以下で，例をあげて，説明をしていく。

上表は，1 行目の 1 セルごとに項目名が列挙され，それ以外の分類情報などは一切無い形なので，項目名－項目値の単純構造にはなっている。実は，原材料名 1 と原材料濃度 1，原材料名 2 と原材料濃度 2 は，2 つの項目がペアになっていることがわかると思う。これが「項目間に論理的関係がある」ということなのだ。焼結温度，焼結風速も項目間に関係はありそうだが，それは性能の高い目的物を作る場合の温度と風速の関係であったり，装置の設計上の制限からくる関係であったりで，論理的な関係性ではないのである。論理的関係性とは，原材料名 x と原材料濃度 x のときに「x が同じモノ同士がペアだ」というような項目名に対して決められた人為的なルールのことである。また，上記には 1,2 を入れ替えても同じ実験になるという対称性もこの項目の論理的関係性として埋め込まれている。ペアや対称性といった項目自体に内在させられた関係性は，物理（自然科学）とは関係のない人為的な，項目名命名による関係性である。そして，データを絞り込んだり，分析をする場合には，この項目の論理的関係からくる論理制約（ペアや対称性）を排除する必要があり，非常に面倒な作業が必要になってくる。例えば，エチレンを使っていない材料を探すにしても上記項目の論理的関係を考慮して，「原材料名 1 にエチレンがなく，かつ，原材料名 2 にエチレンがない，および原材料名 1 にエチレンがある場合は，原材料濃度 1 が 0，および原材料名 2 にエチレンがある場合は，原材料濃度 2 が 0」という条件で絞り込む必要がある。「項目間の論理的関係」が無いように項目名を定義しておけば，こういう面倒さはなくなる。また，「引張強度のプロピレン濃度依存性」を確認しようとしても，この表形式のままでは X-Y プロットグラフを描くことはできない。

以下の表は，上表から「項目間の論理的関係」を排除した項目名である。

この場合は，エチレンを使っていない材料を探す場合，「エチレン濃度が 0」という条件で絞り込むだけでよくなり，「エチレンを使っていない」をそのままストレートに条件にすれば良くなる。また，プロピレン濃度の列を X 軸に設定し，引張強度の列を Y 軸に設定するだけで，「引張強度のプロピレン濃度依存性」の X-Y プロットグラフも簡単に描くことができる。実は，ペアや対称性といった項目名に内在させられた関係性を持つ項目というのは，本来 1 つの項目であるべきものを 2 つの項目に分けてしまったから発生したものなのである。

それでは，原材料名 1 と原材料濃度 1 という項目名は，データ分析が行い難いにもかかわらず，なぜ，よく使われているのだろうか？上記例では，データ記録のための列数は「原材料名 1 と原材料濃度 1」型の方が多くなるが，原材料種類が 100 種類になれば，「エチレン濃度」型は，100 列にもなってしまい，100 列の中で実験毎に使われるのは 2 列だけで，その 2 列を 100 列から探さなければならなく，非常に入力しにくい表形式になってしまう。つまり，研究者はそういうことを先取りして，入力のしやすい形式を選んでいるのである。しかし，この入力しやすい形式が分析をし難くしてしまっているのである。すべてのケースがとは言わないが，入力のしやすさとデータを探したり，分析をするしやすさは，相反関係になる。

実験 ID	エチレン濃度	プロピレン濃度	ブタン濃度	引張強度
EXP1	80	20	0	10
EXP2	25	0	75	12
EXP3	0	60	40	8

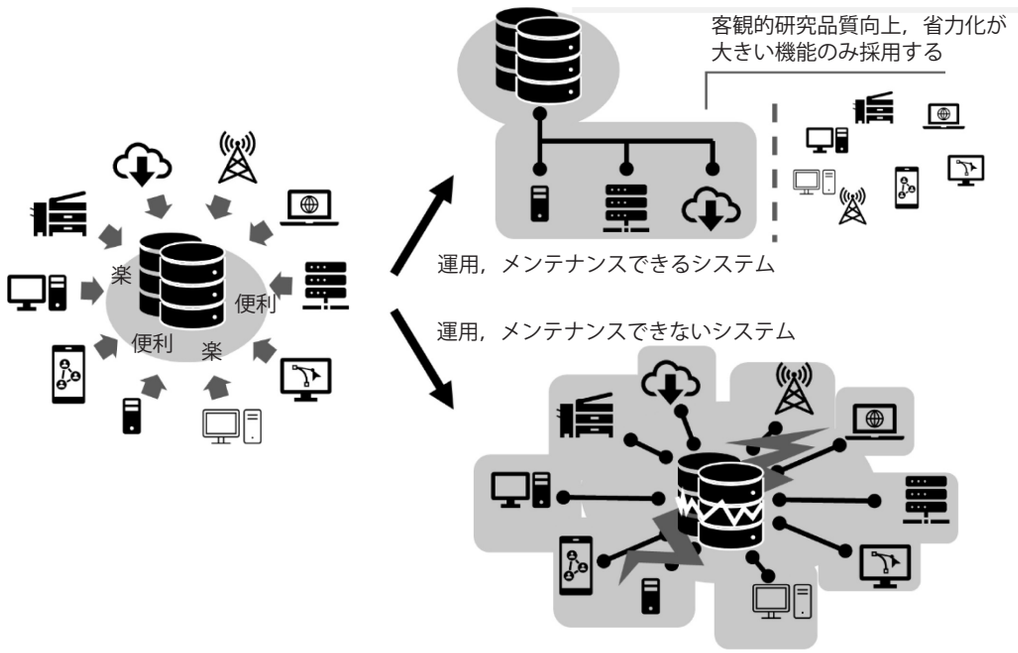

図 23　便利と楽を追及すると運用，メンテナンスできないシステムになってしまう

1 つの Excel ファイルでデータを入力するシートとデータ絞り込み，分析するためのシートを分ければ入力と分析のしやすさの両立は可能である。しかしながら，入力シートと分析シートは参照式や結構面倒なマクロを書く必要があり，どの項目とどの項目がペアだったり，対称性があったりを自動判断させるためには，項目名の命名規則を相当厳密に決める必要もある。項目数が膨大で，かつ項目の追加，変更の多い R & D では，このような参照式やマクロ，項目名ルールを維持し続けることは，ほぼ不可能である。

結局は，R & D ではデータ記録のための入力のしやすさは犠牲にするしかないという結論になる。実際は，入力し難いといったところで，数割程度入力にかかる時間が増えるだけである。その数割の手間を惜しんでしまうことで，データを絞り込み，分析をする毎に入力時よりはるかに多くの手間が発生することになる。当然，入力時の手間を許容できないようであれば，データを絞り込み，分析をする時の手間も許容できようはずがなく，結局，データは使われなくなるのである。

12　項目間に論理的関係のある項目は作ってはいけない

前章では，「記録するデータの項目名は，意味が分かれば何でも良いという訳ではない」と題し，記録するデータの項目名は，どのように決めるべきかを説明した。本章では，「項目間に論理的関係のある項目は作ってはいけない」と題し，項目間に論理的関係のある項目名を作ってはいけない理由をデータ分析の観点から論じる。

以下表の原材料名と原材料濃度は，論理的関係のある項目である。

実験 ID	原材料名 1	原材料濃度 1	原材料名 2	原材料濃度 2	合成温度	引張強度
EXP1	エチレン	10	プロピレン	30	100	350
EXP2	プロピレン	30	ブタン	20	200	710
EXP3	ブタン	50	エチレン	10	300	550
EXP4	エチレン	20	プロピレン	40	400	400
EXP5	プロピレン	60	ブタン	20	500	260

例えば，エチレンの濃度が引張強度にどのような影響を与えるかを調べたいと思ったときに，エチレンの濃度をX軸，引張強度をY軸として，グラフを描きたいと思っても上記表では，非常に面倒だということはわかると思う。また，多変量解析の基礎である重回帰分析をしようと思っても，上記表のままではいろいろ不都合が生じる。以下，重回帰分析を例にして，どのような点で困るのかを説明していく。

重回帰分析は，原則的には項目値が数値である必要がある。重回帰分析が，y を目的変数，xi を説明変数とした場合に，y = f（x1，x2，x3，x4，・・・），例えば y = ax1 + bx2 + cx3 + dx4 + e（a, b, c, d, e は定数）の形の数式になることを考えると，文字列が入っていると演算ができないので，当然といえば当然である。ただし，世の中のデータ分析をしたいものには，文字列が入るもの（例えば，性別，出身都道府県，業種など）も多い。実はそういう時のために，ダミー変数化という処理を行うことで，項目値が文字列の項目でも重回帰分析が可能なように拡張できる。ダミー変数化処理とは，以下のように項目値が文字列の項目をその項目値を項目名とし，値が 0 もしくは 1 となるように変形することである。

ダミー変数化処理を行うとすべての項目は数値化され，重回帰分析が可能になる。上記表では，目的変数が引張強度で，説明変数がエチレン 1，プロピレン 1，・・・・温度の 9 変数になる。機械的に行うと上記表のようになるが，原材料 1,2 が交換可能な対称性を持つという項目間の論理的関係性を考慮すると，以下のように表を変形することも可能である。こうすると説明変数の数は 6 個になる。

実験 ID	エチレン 1	プロピレン 1	ブタン 1	原材料濃度 1
EXP1	1	0	0	10
EXP2	0	1	0	30
EXP3	0	0	1	50
EXP4	1	0	0	20
EXP5	0	1	0	60

エチレン 2	プロピレン 2	ブタン 2	原材料濃度 2	合成温度	引張強度
0	1	0	30	100	350
0	0	1	20	200	710
1	0	0	10	300	550
0	1	0	40	400	400
0	0	1	20	500	260

⇩

実験 ID	エチレン	プロピレン	ブタン	原材料濃度 1	原材料濃度 2	合成温度	引張強度
EXP1	1	1	0	10	30	100	350
EXP2	0	1	1	30	20	200	710
EXP3	1	0	1	10	50	300	550
EXP4	1	1	0	20	40	400	400
EXP5	0	1	1	60	20	500	260

これは，（原材料名 1，原材料濃度 1）と（原材料名 2，原材料濃度 2）が交換可能な対称性をもった変数ペアだからできることであり，どんな場合でもこのように変形していいという訳ではない。つまり，データ表を純粋なデータアナリストやデータサイエンティストに渡しただけでは，このような変形さえできないことを理解しておくべきである。

実際には交換対称性を考慮した上表の 6 個の説明変数でも実は，説明変数としては過剰である。後で説明するが，過剰というのはまだ，考慮できていない項目間の論理的関係性があるということである。一般的に，説明変数の数は，その実験の独立変数の数＝実験パラメータの種類数と一致している必要があり，それよりも多くても，少なくても正しい重回帰分析ができない。正しいデータを分析するためには，項目間の論理的関係性を完全に取り除いた以下のような項目にする必要があるのである。

下表では，説明変数の数が 4 個になっており，これが真の実験パラメータであり，この実験の独立変数でもある。上表では，変数が 4 個なので，1 次式（y ＝ ax1 ＋ bx2 ＋ cx3 ＋ dx4 ＋ e）を想定すれば，5 個の未定係数を決めることができる。ちなみにこれを解くと，以下のようになる。

引張強度
＝ 10 ×エチレン濃度－ 5 ×プロピレン濃度＋ 3 ×ブタン濃度＋ 4 ×合成温度

本当の重回帰分析では，誤差も考慮する必要があるので，未定係数の数の十数倍の実験結果が必要であるが，原理的にはこのようにして，目的変数の説明変数依存性を明らかにすることができる。しかし，項目間に論理的関係があると，本当の実験パラメータではない，見せかけの変数が実験パラメータのように扱われてしまい，いくつかの問題により，上記，目的変数の説明変数依存性は導出できなくなってしまう。

一つ目の問題は，項目間に強い相関があると正しく未定係数が決定できないという重回帰分析の性質である。これは，論理的相関でなくとも，結果論として相関が強い場合も問題となるというものである。例えば，年齢，体重，身長などを説明変数とした分析をする場合，体重と身長には，論理的な相関はあり得ないが，実質的に身長の高い人の方が体重は重いという大きな相関があれば，データ分析が正しくできないというもので，多重共線性問題と呼ばれているデータ分析における基礎中の基礎の問題である。当然，論理的相関というものは，100％の相関であり，重回帰分析において，項目間の論理的相関は絶対に排除しないといけない問題である。

もう一つの問題は，項目間に論理的相関がある説明変数は，独立変数，つまり，真の実験パラメータではないため，どの様な数式を対象数式（モデル式）として採用すべきかが，決めにくくなったり，別途制約条件を設定しなければならないという問題である。

例えば，

引張強度
＝ 10 ×エチレン濃度－ 5 ×プロピレン濃度＋ 3 ×ブタン濃度＋ 4 ×合成濃度

を，説明変数がエチレン 1，プロピレン 1，・・・・温度の 9 変数になる項目名で表し直してみよう。

引張強度
＝ 10 ×（エチレン 1 ×濃度 1 ＋エチレン 2 ×濃度 2）－ 5 ×（プロピレン 1 ×濃度 1 ＋プロピレン 2 ×濃度 2）＋ 3 ×（ブタン 1 ×濃度 1 ＋ブタン 2 ×濃度 2）＋ 4 ×合成濃度

見てわかる通り，エチレン 1 ×濃度 1 のように項目間の掛け算が入ってくる。つまり，y ＝ ax1 ＋ bx2 ＋ cx3 ＋ dx4 ＋ e のような 1 次式を仮定して，重回帰分析を行うのでは，この解にたどり着かないのである。2 次の重回帰分析であれば，エチレン 1 ×濃度 1 のよ

実験 ID	エチレン濃度	プロピレン濃度	ブタン濃度	合成温度	引張強度
EXP1	10	30	0	100	350
EXP2	0	30	20	200	710
EXP3	10	0	50	300	550
EXP4	20	40	0	400	400
EXP5	0	60	20	500	260

うな項は，入ってくるが未定係数の数は 55 にも膨れ上がる。その 10 倍程度なければ，統計的に有意な未定係数が算出できないとすると 550 実験程度の実験が必要ということになる。また，その時に 1 次の項である単独のエチレン 1 や濃度 1 は，係数は厳密に 0 になるべきだが，重回帰分析で係数が偶然にも 0 になることはまずない。明確な形で，「1 次の項である単独のエチレン 1 や濃度 1 の項の係数は厳密に 0 になる」という制約条件を課す必要があるのである。

実際には，9 変数から 6 変数にしたような論理的関係をしっかり考え，項目間に論理的関係が全くないような項目にしてから重回帰分析をすれば，見せかけの変数は消失し，正しい分析は可能になる。しかし，一旦，項目を決めてしまうとそれは結構難しい作業になり，データ分析の時にすぐに漏れなく思い出せるものではない。実際，9 変数から 6 変数にするときは，（原材料名 1，原材料濃度 1）と（原材料名 2，原材料濃度 2）が交換可能な対称性をもった変数ペアということを考慮することで，見せかけの変数を 3 つ削減できたが，真の実験パラメータ自由度＝独立変数の数は 4 であり，さらに 2 つの変数削減が必要である。しかし，どのような項目間に論理的関係があり，どのように変数削減を行うべきかを明確化するのは非常に難しいのではないだろうか？もし，漏れなく変数削減を明確化できたとしても，データ分析毎に項目の変換が必要となり，データ分析が非常に煩雑になってしまう。そもそも，「データ分析の時にどのような変数削減を行うべきかを明確化できる」ぐらいであれば，最初から項目間に論理的関係がないような項目で実験データ記録をしておくべきであろう。

実は，データを分析し易い項目名とは，実験の独立変数を意識した項目のことであり，実験の独立変数に根差した項目はおのずとデータ分析がし易くなる。当然のことだが，項目間に論理的関係があるということは，その項目は独立変数にはなりえないので，何が独立変数なのかということを意識してデータ記録をすればいいわけである。実際，慣れてしまえば，実験の独立変数に根差した項目はデータ入力としても，違和感はないはずなので，データ共有を目指すのであれば，実験の独立変数とは何かをもう一度見つめ直して欲しい。

13 記録するデータの項目名には，分析で使う単位をつけるべき

前章では，「項目間に論理的関係のある項目を作ってはいけない」と題し，項目間に論理的関係のある項目を作ってはいけない理由と回避方法を説明した。本章では，「記録するデータの項目名には，分析で使う単位をつけるべき」と題し，記録するデータの項目名につける単位はどうあるべきかについて，論じる。

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だけの単純な工程を考える。項目間に論理的関係のない一番単純な項目名は，次のようなものになるだろう。

a）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）

この項目名で，合成物の粘度の触媒量依存性を調べる場合，どうすればいいか？ X 軸に触媒重量（g），Y 軸に粘度（Pa・s）を設定して，X-Y プロットを描けばいいだけなのでは？と思うかもしれない。実は，データ分析をするためにはそれだけではダメで，X 軸以外の実験パラメータ，つまり，エチレン重量（g），プロピレン重量（g），ブタン重量（g）のそれぞれが同じ値のデータごとにデータを分類して，X-Y プロットを別グラフとして描く必要がある。そうしないと，X-Y プロットに触媒重量（g）以外の依存性が紛れ込んでしまい，正しい触媒重量（g）依存性をみることができないのである。

ただ，それだけでは「エチレン重量（g）＝プロピレン重量（g）＝ブタン重量（g）＝ 100g」の実験と「エチレン重量（g）＝プロピレン重量（g）＝ブタン重量（g）＝ 300g」の実験は，別グラフのプロットになってしまう。しかし，これらはエチレン，プロピレン，ブタンの比率は同じなので，量効果（容器壁面，表面効果など）が無視できると仮定するなら，粘度など生成物の物理特性は同じになるはずであり，同じグラフに実験結果をプロットしたいはずである。もちろん，エチレン，プロピレン，ブタンの比率が同じものを何らかの方法で分類をして，それを 1 つのグラフにプロットすればいいのだが，データ数や項目数が多くなると非常に手間のかかる作業であり，その作業内で間違いを起こしてしまうとデータ分析が台無しになってしまう。実際に，実験データの記録は間違っていなくとも，その後のデータ処理時の変数値の四則演算などで間違いを起こし，データ

分析を何度もやり直した経験のある方も多いのではないだろうか？現場の研究者はあまり意識していないことが多いが，データ分析時に煩雑な手動処理を避けることは，データ分析の再現性を確保し，データ分析の信頼性を高めるためには実は非常に重要なことなのである。

容器壁面，表面効果などの量効果はほとんど効かないことが多く，量効果を分離した分析をすることが多い場合は，総重量を分母にした重量百分率の単位の項目が便利であろう。

b）エチレン濃度（wt％），プロピレン濃度（wt％），ブタン濃度（wt%），触媒濃度（wt%），総重量（g），粘度（Pa・s）

このような単位の項目にすると，X 軸に総重量（g）を設定し，エチレン濃度（wt%），プロピレン濃度（wt%），ブタン濃度（wt%），触媒濃度（wt%）がそれぞれ一定値のデータで，グラフを描くと総重量依存性が簡単に確認でき，便利である。そして，エチレン濃度（wt%），プロピレン濃度（wt%），ブタン濃度（wt%）のそれぞれが同じ値のデータごとにグラフを描くとすると「エチレン重量（g）＝プロピレン重量（g）＝ブタン重量（g）＝ 100g」と「エチレン重量（g）＝プロピレン重量（g）＝ブタン重量（g）＝ 300g」は，同一グラフにプロットすることができ，a）での問題は解消される。

しかし，よく考えるとエチレン濃度（wt%）＋プロピレン濃度（wt％）＋ブタン濃度（wt％）＋触媒濃度（wt%）＝ 100 であるので，エチレン濃度（wt%），プロピレン濃度（wt%），ブタン濃度（wt%）のそれぞれが同じ値（例えば，それぞれ 15,40,42 wt%）だとすると，触媒濃度（wt%）の値は 1 つの値（3 wt%）しかとらなくなってしまい，X 軸に触媒濃度を設定し，触媒濃度（wt%）依存性を確認しようとしても，プロットは縦に並ぶだけ（触媒濃度＝ X 値は，1 つの値だけ）になってしまう。ちなみに，重回帰分析を行う場合は，濃度のどれか一つ（例えば，エチレン濃度）を変数から外して，つまり，一つの変数を従属変数とし，独立変数の数に合わせて分析すれば，モデル式が適切かどうかということ以外は，問題は発生しない。あくまで，X-Y プロットを描いて，データ分析をしようとするときに問題が出るということである。

次に「触媒濃度（wt%）依存性を確認しようとしても，プロットは縦に並ぶだけになってしまう」問題の解消を試みてみよう。エチレン＋プロピレン＋ブタンの重量を分母にした重量百分率を’wt%と書くこととすると，以下のような項目名が定義できる。

c）エチレン濃度（‘wt％），プロピレン濃度（‘wt％），ブタン濃度（‘wt%），触媒濃度（‘wt%），総重量（g），粘度（Pa・s）

このような単位の項目にすると，エチレン濃度（‘wt%）＋プロピレン濃度（‘wt％）＋ブタン濃度（‘wt％）＝ 100 であり，触媒濃度（‘wt%）はそれらと完全に独立になるので，X 軸に触媒濃度を設定し，触媒濃度（wt%）依存性を確認する場合でも，b）で問題となったようにプロットは縦に並ぶだけになってしまうことはない。もちろん，重回帰分析を行う場合は，エチレン，プロピレン，ブタン濃度のどれか一つ（例えば，エチレン濃度）を変数から外して，分析する必要はある。

本章で伝えたかったことを纏めると，記録蓄積するデータの項目およびその単位は，実験時の記録の取りやすさではなく，実験の独立変数を意識し，さらにデータをどのように分析するのかを具体的に考え，分析に沿った単位でなければならないということである。

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」8 月号，技術情報協会，(2020)，p33-37

5）上島豊，月刊「研究開発リーダー」9 月号，技術情報協会，(2020)，p53-57

6）上島豊，月刊「研究開発リーダー」1 月号，技術情報協会，(2022)，p58-63

7）上島豊，月刊「研究開発リーダー」2 月号，技術情報協会，(2022)，p46-50

8）上島豊，月刊「研究開発リーダー」3 月号，技術情報協会，(2022)，p62-65

9）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
10）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
11）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
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
ケモインフォマティ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