

獨協大学での個人用衛星データ解析システム

Satellite Data Analysis System for Personal Use at Dokkyo University

中村健治^{*1}

Kenji Nakamura

Email: k13002@dokkyo.ac.jp

キーワード：衛星データ解析、Linux、年配者向けシステム

Keywords : Satellite Data Analysis, Linux, System for Senior

獨協大学において衛星の地球観測データの解析を行っている。獨協大学はいわゆる文系の大学でありまた教育を主たる目的とした大学であり、理系の研究大学とは環境が大きく異なる。このような中で大きな容量を持つ衛星データの解析システムを個人用に構築した。このシステムはデータサーバーと解析用 PC から成る。データサーバーの総容量は 140TB に及ぶ。また解析用 PC は 4 式ある。このような大容量のデータサーバーを個人用に使うことのメリット、デメリットを記すとともに、少子化時代を迎えた我が国で今後の年配者のためのデータ解析システムの在り方を考える。

The data analyzing system at Dokkyo University is described. Dokkyo University consists of Language School, Law School, Economics School, etc. and scientific data analysis at Dokkyo University is not popular. In this circumstance, a personal data system is prepared. This system is for analysis of satellite Earth observation data. The amount of satellite data is huge, but rather homogeneous. The analyses include developing prototype data processing algorithms, automatic sequential data processing, and iterations. Since the system is for personal use, processing of huge data is easy to perform without being annoyed by other processes. Data structure can flexibly be designed. On the other hand, the system maintenance, updating the system, installing of new software, etc. should also be done personally. Japan is facing to an aging society, and the number of senior data analysis people could increase. Senior friendly data analyzing systems are also discussed.

*1：獨協大学 経済学部

1. はじめに

獨協大学において衛星の地球観測データの解析を行っている。獨協大学はいわゆる文系の教育を主たる目的とした大学であり、理系の研究大学とは環境が大きく異なる。普通の理工系の大学では、研究室を構え、計算機システムの維持運用、データの入力などはほとんどすべて技術職員、ポスドク、院生が行うことが通常であり、年配の教授レベルでは直接システム構築に携わることは余り無い。しかし、データ解析は研究の現場であり、年配の研究者であってもデータに直接触れていることは重要である。

年配の研究者がいわゆる文系の大学に異動した場合にはデータ解析のために新たにシステムを構築する必要に迫られる。また多人数の研究室形態ではなく一人だけの研究室となる傾向がある。この場合、基本的にはすべてを自分で行わなくてはならない。

本稿では獨協大学という文系の大学において筆者自身が構築したデータ解析システムを紹介する。また、その使用法、ほとんど1人で使うシステムであることの特長性、さらに同業者の解析システムとの比較について述べる。さらに、少子化の下にある我が国で、年配者のためのデータ解析システムの在り方について考える。

2. データ

筆者のシステムに現在蓄積されている、あるいは今後増えていくデータは衛星による地球観測データとそれに関係した地上実験のデータである。筆者が主に使用する衛星データは1997年11月に我が国の種子島宇宙センターからH2ロケットにより打ち上げられた熱帯降雨観測衛星 (TRMM: Tropical Rainfall Measuring Mission) と2014年2月末に打ち上げられた全球降水観測計画 (GPM: Global Precipitation Measurement) の主衛星のデータである。TRMMは日米対等の共同計画の衛星であり、17年以上稼働したが、軌道保持用の燃料が尽きたため運用が停止され2015年5月には地球に落下し燃え尽きた。TRMMが名前の通りに熱帯域の降水の観測を主としていたのに対しGPMは全球の降水観測を目標としており、国際協同による一つの主衛星と複数の副衛星から成る計画である。データは宇宙機関により品質管理された後に配布されている。必要物理量の導出アルゴリズムは数年毎に改訂されていくが、改訂に伴い過去のデータが再処理される。TRMMのデータは現在 version 7 である。GPMのデータは Version 03 である。データ入手はTRMM衛星の打ち上げ後数年は Exabyte テープなどでその後はハードディスク送付に依っていた。現在はインターネット経由での download により直接にデータサーバーに取り込んでいる。転送速度は夜から明け方は十数 MB/s 程度と速いが working day の日中は数 MB/s 程度と遅い。データの種類によるが1種類のデータの一月分の download に長い場合には8時間程度かかる。Working day の日中は大学の回線を遅くしてしまう可能性があり、転送は避けるようにしている。

扱っている衛星データの大部分はTRMM また GPM 主衛

星に搭載されている我が国が開発した降水レーダのデータである。TRMMの観測の概念を図1に示す。主な降雨測器としてはレーダ (PR)、マイクロ波放射計 (TMI) と可視・赤外放射計 (VIRS) がある。レーダは地表面を衛星進行方向に対して直角方向にレーダビームを振ることで面的なデータを得ている。データのレベルには測定生のデータに近いレベル1、必要物理量に変換されたレベル2のデータがある⁽¹⁾⁽²⁾。実際にはこの下流にさらに時空間平均された高次のデータがあるがデータサイズは小さくなる。筆者がもっともよく使用するデータはレベル2のデータである。TRMMの降雨レーダの走査幅は約200km、水平分解能は約4kmである。またレーダであるので距離分解能があり、250m毎 (レンジビンと呼ぶ) に地表面から高度20km弱までを観測している。一つのビーム (ray) では80個のデータがあり、一つの走査は50個のrayから成っている。地球一周では約1万走査あるので、地球一周のデータ (1 granule と呼んでいる) では40M個のデータがある。4byteデータとして約160Mbytesとなる。レベル2のデータは、いくつか種類があるが、もっともよく使われるデータは一つのrayのレンジビン毎に二つデータがあるのでさらに大きくなり330MB/granule程度となる。現在運用中のGPM主衛星の降水レーダは二周波レーダであるので複合データも含めてデータ量はさらに増え約750MB/granuleとなっている。衛星は低軌道衛星であるので地球一周にかかる時間はほぼ1時間半である。1日では約16周となり、結果としてTRMMのレーダについては5GB/日、GPM主衛星のレーダについては12GB/日のデータ容量となる。TRMMのデータは通常は圧縮されている。圧縮により20分の1程度になるが、データ処理では解凍のためにかなりの時間がかかる。他のマイクロ波放射計などのデータは2次元データなので、レーダのデータほどには大きくはない。

衛星データ以外にも地上での衛星データの検証実験のデータも扱っている。データは500GBから1TBのポータブルハードディスクで受け取っている。また気温、湿度、風、雨滴粒径分布などの現地気象観測データもあるが、これらは容量が小さく、また随時入ってくるため常時使用しているMS WindowsのノートPCで受け取り、データサーバーに転送している。現在のデータストレージ状況を下に記す。

地上レーダ: 10TB (一部bz2圧縮)
過去気象データ等: 20TB
TRMM: 2TB (gz圧縮)
GPM: 10TB

今後数年間では50Tの増加が予想され、数年後には100TBに近づくが、現在のデータサーバーのストレージ容量は140TBなので十分に賅える容量である。

3. 解析システム

上のような大きなデータを扱うため大きな容量を持つ衛星データの解析システムを個人用に構築した。このシステムはデータサーバーと解析用PCから成る (図

2)。データサーバーは2式あり総容量140TBになる。また解析用PCはサーバー内の2式を含めて4式ある。なおデータの容量は大きいがいわゆるビッグデータではない。世にいうビッグデータは様々なやり方で収集される雑多なかつ大量のデータを指している。このようなデータから有用な情報を抽出することは確かに今日的課題であり、必要とされている技術であるが、筆者の扱うデータは種類の少ない画一的なデータであり、サイズは大きいがビッグデータとは言えない。

解析用PCはLinux PCである。これに加えて筆者が通常使っているWindows ノートPCにWindows上で動くUNIX likeなCygwinを入れて使っている。Cygwinは自宅での作業用である。サーバー (calc0、calc1) とフロントエンドの一つ (NPC : New PC) は学内の研究所棟内に置いてありLocal LANの固定IPアドレスを持っている。もう一つのフロントエンドPCとして筆者の居室にNewOPC (New Office PC) を持っている。IPアドレスはDHCPによる割当てである。NewOPCにはlocalでEPSONのインクジェットプリンタを接続してある。研究所棟内部は1GのLANとなっているが研究所棟の外部への回線は100Mである。学外へは近く300Mに増強される予定である。サーバーはメーカー (HPCテクノロジー社) 製であり、フロントエンドの2式は組立てPCである。サーバーは部屋を仕切って小さい小部屋を作りそこに設置してある。この設置では新たな空調の設置、天井の火災報知器の移動、またサーバーに負荷がかかったときは意外に大きな音がするので騒音防止のための防音壁の付加が必要であった。

各マシンのCPUは

calc0、calc1 : Intel Xeon、E5-2620、2.00GHz
NPC : Intel Core i7-3770K、3.50GHz
NewOPC : Intel Core i7-4770、3.40GHz

となっている。Intel Coreの最新版はCore i7-5xxxなので若干古くなっている。メモリ容量は

calc0、calc1 : 16G、NPC : 32G、NewOPC : 8G

となっている。NewOPCのメモリ容量は若干小さいが不便を感じたことは無い。ストレージはNPCとNewOPCはSSDとハードディスクを組み合わせている。

Linuxのversionは4つのPCとも2.6.32である。最新は3.xなので若干古い。OSはすべてRPM系RHELクローンで安定性に定評のあるCentOSである。Versionは

calc0:6.2、calc1 : 6.4、NPC : 6.5、NewOPC : 6.5

である。CentOSの最新版は7.xなのでこれも若干古くなっている。すべてのPCでOSはほぼ同じであるが、細かくは異なる。例えばデータサーバーであるcalc0、calc1ではdefaultのshellはC-shellであるが、筆者の通常のshellはbashでありNPCとNewOPC上ではbashを使っている。変更は簡単であるがデータサーバーについてはなるべく構築した時の形を保持するようにし

ているのでそのままとしている。フロントエンドのPCのOSには、現在もっとも普及していると思われるUbuntuの使用も試みたが、MS Windowの代替えを狙っているようでプログラム開発向きではないように考えたためCentOSとしている。

すべて64ビット型としている。このためどのPCでcompile-linkをしても同じ実行形式のコードが他のPC上で走る。

実際に処理を行うと単独処理ではサーバーのcalc0、calc1よりもNPCとNewOPCの方が速い。しかしCPUがXeonであるcalc0、calc1は10個近くの並列処理ならばcommand promptの反応は悪くなるものの処理速度そのものはあまり劣化しないようである。

4台のPCすべて非常に安定している。サーバーは増設時と定期点検以外は常時稼働しているようにすべて電源を入れたまま何か月も置いているが、ソフトウェア上のfreezeは生じていない。ハードウェア上では、電源系と思われるトラブルでNPCが落ちたことと、サーバー上のRAIDシステムの中のハードディスクが1台不具合を起こしたことがあった。後者ではホットスワップで交換ができてシステムの停止には至っていない。

解析用PCが4式あると、データサーバーとフロントエンドの差を一応は設けてはいるが混乱し易い。プログラム開発でも居室のNewOPCに21インチのdisplayを二つつないで並べて画面を広くして、それぞれのPCからのX windowをPC毎に色を変えて表示して使っている。また二つのサーバーではRAIDストレージを相互マウントしているので、その中に開発・処理プログラムを入れたdirectoryを作っている。サーバー上でもプログラム開発を行うが、開発はホームdirectoryの下で行っている。プログラム開発もあちこちのPC上で行うとすぐに混乱するので、一旦開発がfreezeしたプログラムは相互マウントしているデータ領域の中の一つのdirectoryに集めるようにはしている。開発履歴はソースコードになるべく書き込むようにしているがどうしても不十分となる。Git^③を使うなどしてrevision管理をすることが必要である。様々なutilitiesもなるべく同じになるようにしているが、データサーバーの2式はメーカーがOSをインストールしているので変更はなるべく避けるようにしている。フロントエンドとして使用している2式では試行的ためのインストールなども自由に行っている。RedHat系であるCentOSでは流通しているプログラムのインストールはxxxを目標プログラム名としてyum install xxxとすればよく非常に簡単である。一昔前のmakeなどによるインストールは不要である。分からなくてもインターネット上で検索すればインストールの情報が得られる。

4. データ解析の概要

データ解析の主な部分は大容量のデータセットから必要な物理量を抽出し、次に他のデータと突き合わせながら分類し、さらにその時間・空間平均をとることである (図3)。これにより降水システムの気候的特徴を研究する。地上レーダのデータ解析ではFFT (Fast

Fourier Transform) 処理を数多く行う必要もある。より高次の解析としては主成分解析に似た EOF

(Empirical Orthogonal Function) 解析などもある。処理過程ではデータセットからの抽出・分類して中間データセットを作成するが、この段階が処理時間としてはもっともかかる部分となっている。また中間データセットの作成も、プロトタイプを作成してチェックはするものの大容量のデータを処理した後での結果を見ないとその評価は困難である。このため、中間データセットは元データセットよりもはるかに小さいが、様々な解析に対応できるだけのデータ内容と柔軟性を持っていなければならない。この辺りのデータセットの設計が重要となる。しかしながら実際は解析を進める中で場当たりに決めていく。大容量のデータサーバーがありまた自動処理により時間はかかるものの昼夜を問わずに自分のためだけに計算機を走らすことができるという環境のために、気兼ねなく大容量データの再処理を行うことができる。これは個人用大容量データサーバーの良い点として挙げられよう。その一方、処理の手順や処理段階などは自分で記録を残しておくなければならない。

扱っているデータとして図4に若干特殊な例を示す。TRMM は寿命末期になり燃料が枯渇し、軌道制御が十分ではなくなり、定常の運用が中止された。この機会に特殊運用実験がなされた。図4の下の図は搭載されたレーダの観測域を示す。途中で幅が狭くなっているが、これは衛星の方向を90度変更してほぼ同じ線上を何度も走査したためである。上の図は降雨の鉛直分布である。スパイク上のものが降雨エコーである。横に帯状に広がっている部分は地表面のエコーである。この図では走査端のデータを示しているため斜めに観測しているため地表面エコーが広がってしまっている。このような図を大量に処理して平均値等を作り出している。図5は1998年から2013年の16年間の6月から8月、地方時午後6時から12時までの降雨時の降雨の高さの分布図を示している。降雨の多いところでは降雨の高さの高い傾向があるが、陸上で高いこと、チベットでは地面標高が高いこともあって特に高いことなどが見て取れる。この図は、一周毎のレーダデータから降雨域を抜き出し、その時の降雨高度を緯度経度0.1度のグリッドに落とし込み、その後、その平均を取ることによって作成している。処理したデータ量は膨大であるが、処理時間などにわずらわされることなく処理を進めることができている。またPCが複数あるので、並列で処理プログラムの開発がなされている。

データサーバーの容量が大きいためデータ容量を気にしなくてもよい。システム容量に気を使う必要が無いだけでなく、不要データの削除、ゴミの廃棄を強制的にやる必要も無い。また大容量のデータは通常は圧縮されているが、この解凍に実際にはかなりの時間がかかる。これを非圧縮で置いておくことは処理時間の大きな削減になる。筆者は解凍するときは元データの破壊を恐れ、作業directoryにコピーしてから解凍しているが、1回解凍してそのまま置いておけばこのような中間作業が無くなる。すべてを一人で行わなければ

ならない状況の中では処理手順が簡素であることは重要である。

プログラム群の構造も重要である。データ解析において様々なテストを行うとプログラムが沢山作られる。テスト中はプログラム群をまとめておく余裕は無いため多数のプログラムが雑然と積み上がってしまう。ガイドラインの一つとして、処理パラメータなどを自分に合わせて自然な形にしておくことがある。コンピュータシステムの様々な設定では使用者が自然に使えるような形にしておく、またデフォルト値も一般的にしておく、という原則がある。個人使用なのでシステムはどうにでもできるが、それでも自然な形にすべきという原則に則るべきである。不自然な形でも、プログラム内でチェックをかけるようにすればよいが、どんどんプログラムを作成・改訂していく中では十分なチェックルーティンを入れることはできない。

PCに高い能力がありまた複数のハードディスクにより数十TBの容量を持たせることも可能な現在、データの入手・処理のやり方も変化があろう。20年近く前に、衛星計画の中でデータシステムの議論があった。その当時は衛星データの量から宇宙開発機関がデータセンターとしての役割とともに、データの切りだし等も行うことが期待された。現在は、データに関しての宇宙機関の役割は、衛星の元データ及び標準プロダクトと呼ばれる測定されたデジタル量から物理量に変換されたものと、それからの高次物理量を含んだデータセットを生産・維持・保管・配布すること、またアルゴリズムの改訂に伴う再処理となっている。地球温暖化など地球の気候システムの変動・変化の研究のためには長期にわたる様々なデータが必要であり、その量は膨大となる。突発的な災害などではスナップショットとしてのデータも必要で、これに対しては場所や時刻を指定してのデータ切出しが必要となり、これらにも対処されている。しかし、研究的解析では各解析者が必要データをすべて入手し手元で大量のデータを解析するスタイルが普通となっている。筆者の解析方法もこの線に乗っている。

5. 衛星データ解析環境

筆者の環境の一般性を調べるため同業者の環境に関して以下の項目について若干の聞き取り調査を行った。

- OS : Windows、UNIX (Linux、Mac-OS、Cygwin、他)、IDL、他
- 解析マシン (メーカー製、メーカー組立てPC、自作組立てPC、Windows PC、他)
- 言語 (Fortran、C、C++、他)
- 画像出力ソフトウェア (GRADS、GMT、他)
- 画像処理ソフトウェアの利用
- 衛星データの読み出しソフトウェア

結果としては、OSはほとんどがLinuxであった。言語はCとFortranが半々、それにIDL⁽⁴⁾が一部である。サーバーはメーカー製あるいは仕様を示しメーカーで組み立てたPCが多い。使用形態は研究室の皆で使う、

という形が多い。システムの保守は学生あるいは技術職員が行っている。学生が保守する場合は博士課程前期課程で出てしまう者が多いこと、全面的に責任をもってもらうわけにはいかないことなどの問題がある。図示化ソフトウェアはそれぞれである。図示化のみはWindows上で動くフリーソフトウェアを使っている人もいる。衛星データはHDFフォーマット(Hierarchical Data Format, <http://www.hdfgroup.org/>)になっているが、読み出しソフトウェアは、主に衛星に特化した読み出しプログラムを使用している人が多い。一部の人はHDFを直接に扱うソフトウェアを、また少数の人は商用ソフトウェアであるIDLを使用している。基本的には筆者のシステムと似ているが、筆者はただ一人で保守しまた使っている点が異なる。これは筆者の環境が少し特殊であることによっている。

6. 大型データサーバー

個人レベルでのデータサーバーの容量が大きくなったが、それ以上に地球システムに関するデータはモデルの出力データ、観測データの客観解析データや衛星データを中心に膨大となっている。モデルは時空間分解能が上がり、グリッドサイズが2倍になれば時間分解能も計算安定性を保つために1/2にしなければならないので2の4乗で大きくなる。衛星データも地球観測衛星の数の増大、測器の多チャンネル化、空間分解能の向上などで膨大となっている。2015年7月には時空間分解能が大きく上がり、またチャンネル数が大幅に増えた新型の静止気象衛星「ひまわり」の運用が始まりデータ量も大量となっている。

このようなデータをアーカイブするためペタバイト以上のストレージ容量を持つサーバーが必要となってきた。我が国ではこれに対応する大型のデータセンターとして、気象関係は気象庁が構築している。衛星データ関係は国立研究開発法人宇宙航空研究開発機構が持っている。モデル出力なども含めた総合・統合データセンターとしては東京大学が主導で構築している地球環境情報統合プログラム(DIAS: Data Integration & Analysis System)がある。国立研究開発法人情報通信研究機構でも大型大量データのハンドリングと配布の技術開発のため大型のデータアーカイブシステムを構築している。また千葉大学リモートセンシング研究センターでも衛星データを中心としたデータセンターを運営している。若干小規模の千葉大学のシステムでもストレージ容量は2PB程度ある。DIASではデータアーカイブとともに、大量のデータの中からのデータ検索・切り出し、解析ができるようなデータのアーカイブ・配布・解析の大規模の統合システムとなっている。規模は全く異なるが二昔前のセンター計算機の形に戻っている面がある。

数多くの大容量のデータを組み合わせた解析をする場合には大型統合システムの上での作業が効率的となる。しかし、大容量ではあるが限られた数のデータを使う場合は、上のような大型のデータサーバーからdownloadして手元のサーバーを利用した解析となることが通常の形であろう。

7. Senior friendly なシステム

日本は急激な少子化の中にあり、年配者が多数を占める社会に向かっている。しかし、年配者といっても健康のレベルが上がり、元気な年配者が多くいる。この中で、社会の活性を保つためには、年配者も活力の利用を考えるべきであろう。

年配者のデータ解析と解析システムを考えるために、年配者の一人である筆者のやり方を振り返ってみる。

筆者のやり方の特徴は、使用者、システム維持・運用者が年配の1個人であることにある。

まず、システムの維持・運用について考える。この作業はデータ解析を主に行っている研究者では一般には困難であろう。データ解析を主とする研究者でも解析システムについてはある程度は知っておく必要はあるが、実験研究者が実験装置を作り、使用するようなレベルでは全くない。筆者は解析システムの構築作業も行っているが、システムの細かいtuningはできず、お仕着せに近いdefaultのシステム構築をしている。

次に、解析プログラムについて考える。C言語やFortranが一般的に使われるが、これらの言語ではすべての処理を細かく書かなければならない。大きなプログラムを書くことは肉体的労働に近い面があり、若者ではなくては厳しいとも言われる。このため年配者ではCやFortranは適当な言語では無い可能性がある。一方、解析結果の評価では年配者は経験が豊富である分、能力が高いとも言える。このため、関数型言語やIDL⁽⁴⁾のようなデータ解析に特化した言語が一般のデータ解析者ばかりでなく、年配者には特に必要と考えられる。IDLは欧米ではかなり普及しているソフトウェアであり、ベクトル計算を簡単に記述できるとともに、汎用の数学ライブラリが組み込まれており簡単に使うことができる。またHDFやNetCDF(Network Common Data Form, [http://www.unidata.](http://www.unidata.ucar.edu/software/netcdf/)

<http://www.unidata.ucar.edu/software/netcdf/>)のような汎用の大量科学データ用のデータフォーマットにも対応している。これは商用であり購入経費がかかる。GDL(GNU IDLの略であろう)というフリーのクローンがあるが、機能が若干落ちているようである。

年配者にとってはキーボード等の操作性も重要な要素となる。マウスは入力装置として一般的に使われるが、テキストベースで作業をしていると、マウスを使う時にキーボードから手を離さなければならないことが負担となる。このことは以前から知られており、ポインティングスティックやタッチパッドが開発されている。また手の負担の少ないキーボードも開発されている。これらは年配者の使用を特に考えているわけではない。しかし、社会全体は高齢者向けの小物が増えてきている現在、PC用の入力装置にも高齢者の使用を特に考えたものがあってもよいであろう。

システム全体の構成も課題となる。複数のPCの組み合わせは、それぞれの役割を明確にする必要がある。筆者の構成を考えると、まず4式のPCは多すぎる。データサーバーはこのままとし、また定常自動処理をデータサーバーで行うことはデータ転送がlocalで済む

ため速いことなどから自然であろう。フロントエンドのPCは一つとして他はバックアップと考えるべきであろう。プログラム開発はフロントエンドのPC一つとし、古くなったときにはそれをバックアップPCとし、新たにPCを導入してフロントエンドPCとすればファイル移動などもスムーズとなろう。また新PCのシステムインストールに躓いてもやり直すだけで済む。実際、新システムのインストールは何度も行うことが多い。プリンタはネットワークプリンタとしてすべてのPCから出力できるようにすべきであろう。Shellも統一したいところである。要はできる限りシームレスにすべきということである。新しいapplicationの導入などはフロントエンドのPCのみで行い、十分に試された後に他のPCに展開するのがよいであろう。

システムの維持についても述べよう。プログラムのインストールには課題がある。衛星データは通常、読み出しプログラムが宇宙機関から配布される。これはwebsiteに行けば最新versionがsourceあるいはbinaryで手に入る。しかしインストールには手の掛かることが多い。これは必要な下位プログラムが事前に必要であること、環境設定が必要であること、などのためである。解析マシンはLinuxとして、開発タイプの標準のLinuxシステムに衛星データの読み出しプログラム、必要な下位プログラム、また衛星とデータに関するマニュアル等をすべて含んだSatellite-Linuxとも呼べるLinuxの配布versionを作ることがよいと思われる。

衛星データの読み出しプログラムはどんどん改訂されている。通常は互換性が保たれているので問題となることは少ないが、時にはversion upしなければならない。MS Windowsもsecurity対策などで改訂プログラムが配布されており、自動設定をしておけばPCは勝手に改訂する。自動version upまでは要求せずとも、yum installで簡単にできるとよい。このやり方は必ずしもseniorのためばかりではない。衛星データの利用を呼び掛けると常にデータの読み出しが難しい、という苦情がでてくる。この苦情への対応策ともなる。

年配の解析者が増えることが予想される中で、社会全体としてはどのように進むであろうか。一つの方向はクラウドコンピューティングかもしれない。データサーバーシステムの保守・維持は会社に任せて個人はそのシステムを利用するだけ、という方向である。これは一つの潮流となっているようであるが、筆者の属する衛星データ解析の分野ではデータ配布にはクラウドは使われ始めているが解析には及んではいない。この場合でもデータ読み出しプログラムのような特殊プログラムは自分でインストールしなければならないだろう。宇宙機関などが自前でデータ解析システムを含んだクラウドを運用することも考えられるが、これは単に宇宙機関の持つ計算機にアカウントを持つ、という1時代前のスタイルである。強力なPCが普及し、個人レベルで自由にシステムを構築・維持・改訂ができ、また自分のデータシステムを構築できるという利点を生かすには当面は”Satellite-Linux”が良いと思われる。

8. おわりに

獨協大学における筆者のデータ解析システムについて述べた。サーバー容量は大きいこれは衛星データそして降雨に関する地上観測データの生データを収集しているためである。それにしても100TBを越える容量のサーバーを個人用として整備できるようなことは以前では考えられなかったことである。このような中でデータ解析では自動化が大きな役割を占める。データ処理のプロトタイプは試行錯誤を繰り返しながらの開発となるが、一旦出来上がったならば、サーバーに入っているCPUを用いて、昼夜を問わない処理を行う、という形が通常の解析となる。この時も再度のやり直し処理も厭う必要は無い。

個人ですべてを管理しなければならない状況はすべてを個人で決定できるという面では簡単であるが、個人がやらなければ何も進まない、ということでもある。この場合、手順が簡素であることは重要である。データもoriginalデータ、中間データ、最終結果、またプログラムについても中間処理用プログラム、出力プログラム、さらに多数の小物のコマンド等をなるべく自然なdirectory構造の中に収めなければならない。この時、開発用のものと大量自動処理用のものと区別もある。

個人ですべてを構築でするということは各個人の趣味、やり方にシステムを合わせるができるので使用者にとっては居心地が良い。その一方directory構造などを趣味に走って作ってしまうと時間が経った後で使おうとすると忘れてしまっており混乱する。このためdirectory構造などは使用の目途が立った時点で少し考えてから作るように心がける必要がある。それでも1人で使う場合にはシステムの「進化」は遅いものとなる。理工学系の研究室で共通で使用するならばdirectory構造なども学生が作り、また使用に合わせてtuningすることでシステムが「進化」し、年がある程度たつと新しいシステムが入り、新たなシステムが「誕生」するであろうが、個人のシステムではそのような形にはなりがたいと考えられる。

このような状況のもとで、年配の解析者のためのデータ解析システムの在り方を考えると、システムの簡素化、適当なプログラミング言語の選択、年配者向けの入力装置の開発などが重要となろう。データシステムの管理・維持・変更はかなりの労力と時間がかかる。また秘密にしなければならないものはほとんど無いとはいえ、サイバー攻撃が多い中でのsecurityの確保も必要である。年配者としての方法としては、

- (1) すべて自分で構築する、
- (2) クラウドですべての解析を行う、
- (3) 宇宙機関などのデータ供給機関のサーバー上で解析を行う、
- (4) 他のグループのシステムを使う、
- (5) ”Satellite-Linux”を使ってシステムを構築する、

などが考えられる。(1)は筆者が行っているやり方である。しかし必ずしも一般的ではないであろう。(2)

は一つの潮流であり、一般にはかなり普及する可能性がある。(3)は少し古い形である。(4)は多くの研究室での形であろう。研究室では学生の中にはPCシステムやネットワークに興味を持つ者がいる可能性が高く、その学生が様々な面倒を見てくれる。(5)は今後宇宙機関などに働きかけたい方法である。

Seniorのデータ解析システムとしては、いかにして手間暇を省くか、ということが重要と言えよう。解析方法を理解し、解析結果を見る目はあるので、解析処理を如何にして楽にするかが要点であろう。

謝辞

本稿で述べたデータシステムは日本学術振興会科学研究補助金基盤(B)24340110により構築した。構築に

当たっては獨協大学の研究教育推進課、施設事業課の協力を得た。データは国立研究開発法人宇宙航空研究開発機構(JAXA)から得ている。また本研究の一部は情報科学研究所研究助成によっている。また査読者からは有益なコメントを頂いた。

参考文献

- (1) 宇宙航空研究開発機構、地球観測データ利用ハンドブック -TRMM編- (2006)。
- (2) 宇宙航空研究開発機構、全球降水観測計画(GPM)データ利用ハンドブック(2015)。
- (3) 川野辺正博、アリスとボブのGit入門レッスン、秀和システム、pp.281(2012)。
- (4) 伊藤ただし、IDLプログラミング、講談社、pp.298(2013)。

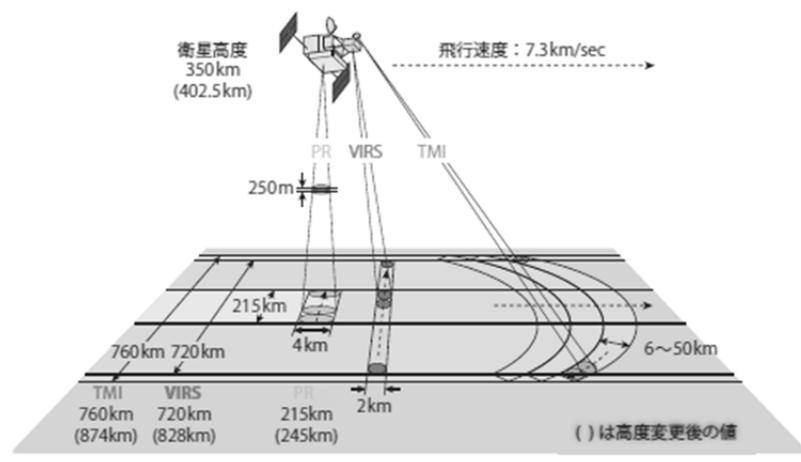


図1 熱帯降雨観測衛星(TRMM)の観測の概念。降雨センサーとしてはレーダー(PR)、マイクロ波放射計(TMI)、可視・赤外放射計(VIRS)がある。PR(レーダ)は衛星進行方向に対して直角面内を走査する(JAXAホームページより)。

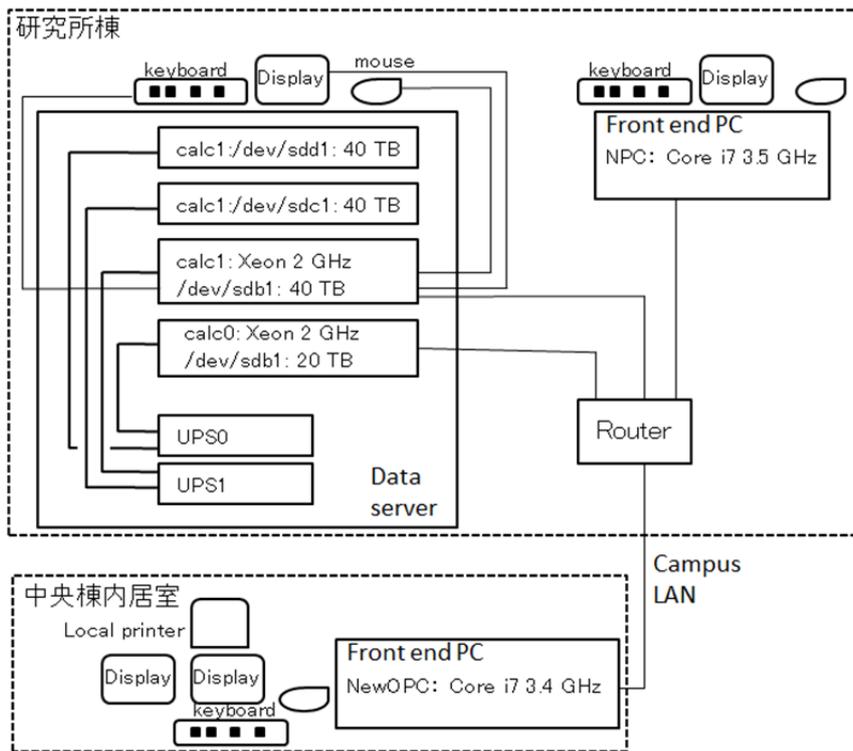


図2 システム図。PCは4式あり、3式は研究所棟に、1式は居室に置いてある。

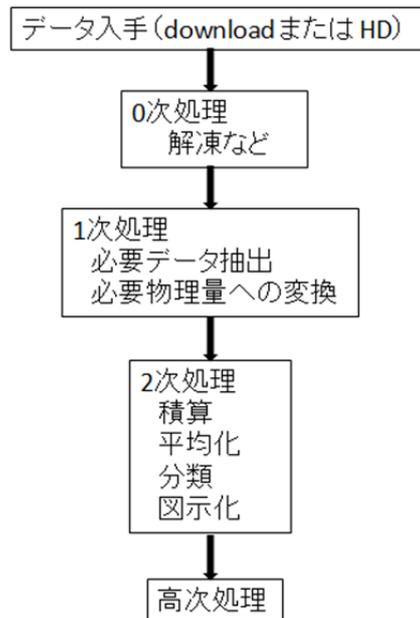


図3 データ処理の流れの概要。

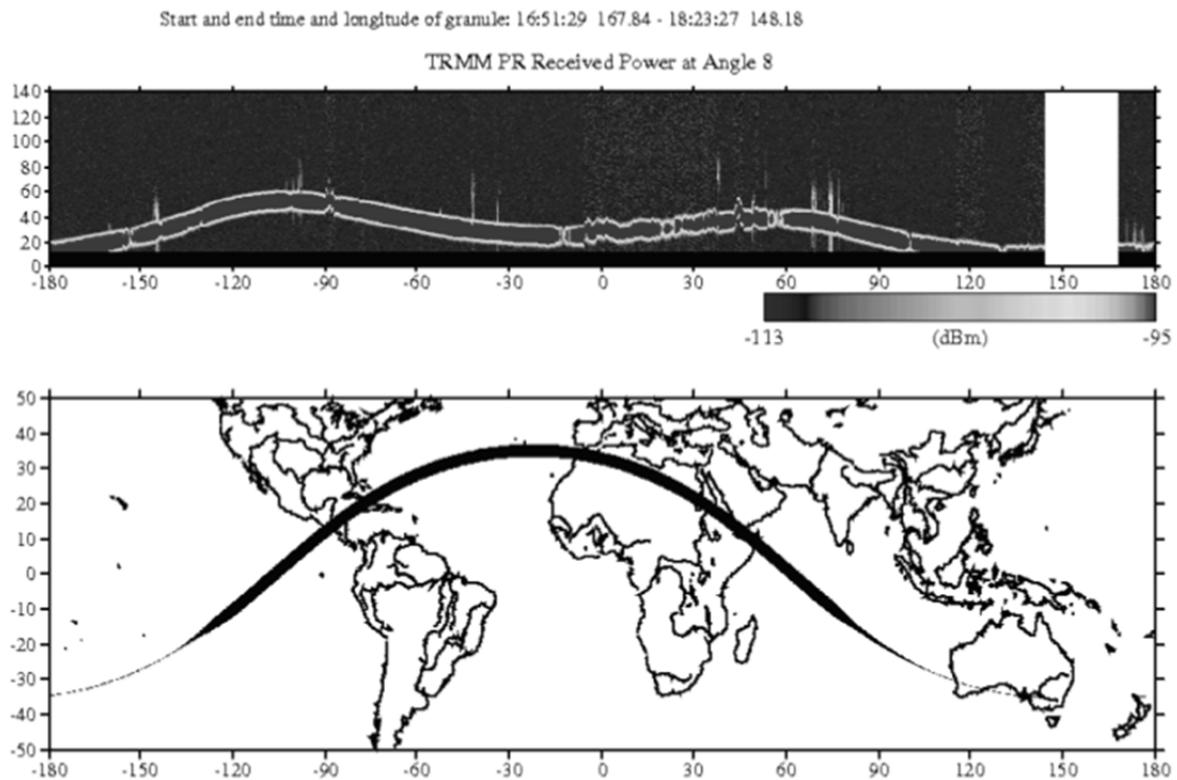


図4 TRMM衛星の運用末期の搭載レーダ特殊観測時の例(2014年11月16日)

(上) レーダのエコー。白い線で囲まれた帯は地表面エコーであり、その上のスパイク上のものが降水エコーである。レーダビームの方向を大きく斜めにしているので地表面エコーが広がっている。(下) 搭載レーダの走査域。両端で細くなっているところが特殊運用の時。レーダの走査を衛星進行方向に直角面内から平行面内に変えている。

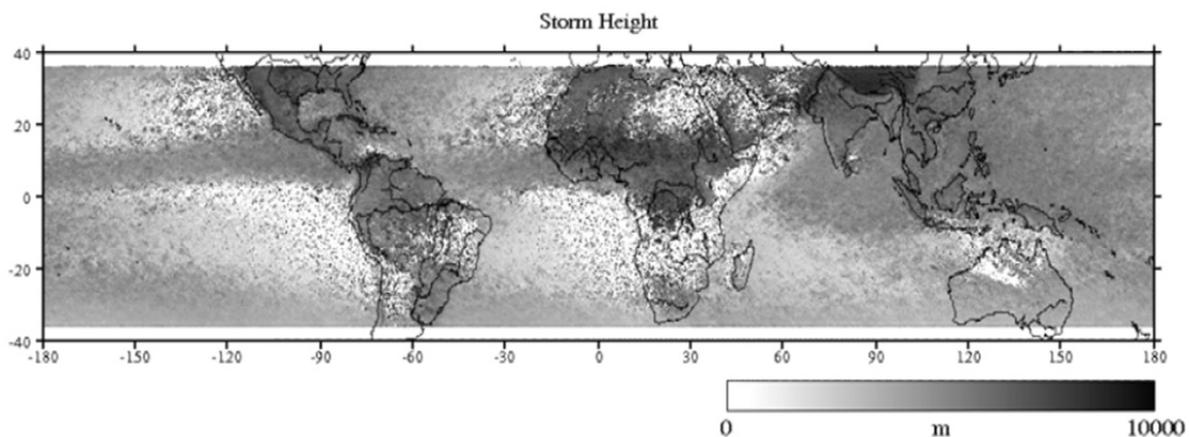


図5 TRMM衛星搭載のレーダのデータを使って作成した1998年から2013年の16年間の6月から8月、地方時午後6時から12時までの降雨時の降雨の高さの分布図。降雨の多いところでは降雨の高さの高い傾向があるが、陸上で高いこと、チベットでは地面標高が高いこともあって特に高いことなどが見て取れる。

(2015年9月29日受付)

(2015年12月2日採録)