ＭＴ法データの圧縮方法について

高知大学理学部地質学教室 村上 英記

Data Compression Method for Magneto-telluric Data

Murakami Hideki
Department of Geology, Faculty of Science, Kochi University

１．はじめに

地球電磁気学分野における時系列観測データは，臨時観測の場合であっても一般に非常に大量なものとなる．観測場所が比較的アクセスのしやすい場所である場合には，記録の交換なども楽であり人手も必要としないが，そうでない場合には記録をどう取るかが観測上のおおきな制約となる．現在のところ，この問題は大容量の記憶装置と電源を使うことで解決されている．しかし，これも観測機器の総重量を増すという問題があり，この方法を取れない場合もある．観測機器にデータ集録だけでなくデータ圧縮の機能があれば，機材の重量を増やすことなく観測期間を延ばすことも可能となる．

また，各観測点で収集されたデータをインターネットなどのネットワークを通じて収集あるいは配布する場合においても，記憶装置やネットワークへの負荷という観点からデータ圧縮は重要な課題となる．汎用のデータ圧縮プログラムを使うだけではなく、データの特性を利用したより圧縮率の高い圧縮技法を開発する必要がある[1),2)]．

データ圧縮では、圧縮率のみを問題とする場合もあれば，計算時間も考慮する必要がある場合もある．ネットワークでデータを転送あるいは蓄積する場合には，圧縮率のみを考慮すればよいが，観測機器にデータ圧縮の機能を持たせる場合には計算時間も消費電力との関係で考慮すべき重要な要素となってくる．

以上のような観点から，時系列データの一例としてMagneto-telluric法で収集される磁場３成分と電場２成分のデータを取り上げ，データ圧縮法を検討した．

２．ＭＴ法データについて

今回の評価に使用したデータは，1991年に実施された琵琶湖西部の総合観測[3)]で得られた三国峠のデータを使用した．使用したデータは，U30により集録されたサンプリングレートが１Hzの磁場３成分と電場２成分のデータである．使用した磁場と電場を図１～図５に示す．

U30では16ビットのADコンバータが使用されており，１サンプルが２バイトのデータとして記憶されているので，圧縮の評価にあたっても１データが２バイトとして扱うことにする．

３．データ圧縮

データ圧縮法には，可逆圧縮法と非可逆圧縮法の２つの圧縮方法がある[4),5)]．可逆圧縮法というのは，圧縮したデータを復元すれば完全に元のデータに戻る圧縮方法であり，非可逆圧縮法は復元したデータが元のデータとは完全に一致はしないが，データの構造を利用して高圧縮率を達成する方法で，画像データなどに利用されている．今回は，可逆圧縮によるデータ圧縮方法についてのみ検討する．

代表的な可逆圧縮法には，Huffman符号化，算術符号化（Arith），動的Huffman符号化（Aduptive Huffman），Lempel-Ziv符号化（LZ-Slide,LZ-Squeeze）などがある．これらのアルゴリズムを組み合わせて汎用圧縮プログラムが開発されている．生データをこれらのアルゴリズムを使って圧縮した場合の圧縮率（＝圧縮されたデータ／生データ×100％）を図６ａに示す．図６ａには，比較のためにLHaというパソコン等で利用されている汎用圧縮プログラムによる結果ものせてある．LHaは，Huffman符号化とLempel-Ziv符号化を組み合わせたものである．図６ａから，磁場データに関してはHuffman符号化，動的Huffman符号化，算術符号化がLempel-Ziv法よりも高い圧縮率を示している．しかし，電場データに関しては逆にLemple-Ziv法の方がよい結果を示している．５成分のデータ全てに圧縮率のよかったのは，LHaであったが，圧縮率は75％～55％程度で

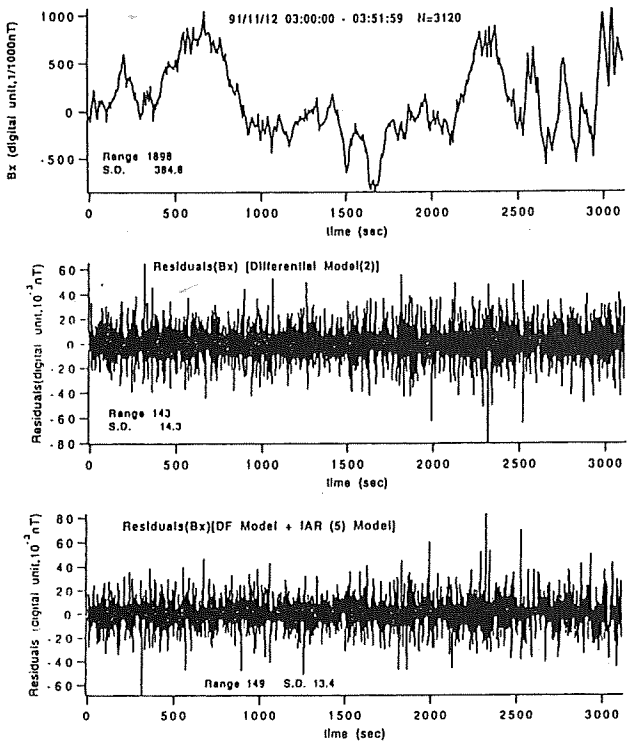

図１．磁場データと残差データ

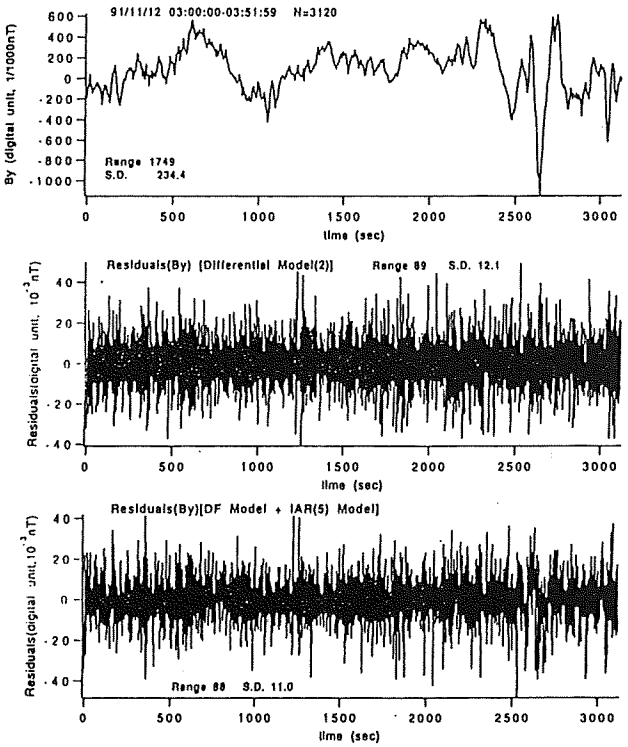

図２．磁場データと残差データ

あった．これらの結果から，データの変動幅と圧縮率が関係していることがわかる．また，Huffman符号化などのデータの頻度分布をもとに圧縮をおこなう圧縮方法と，Lempel-Ziv符号化のような前後のデータの関連性を使う圧縮方法では，磁場と電場を圧縮する場合に違いがみられることがわかる．これは，検討に利用した磁場データは長周期成分が卓越しているために，比較的短区間のデータの前後関係を利用するLempel-Ziv符号化ではあまり圧縮率があがらないことを反映している．同じことは，電場の2成分についても言え，長周期成分の卓越したEx成分の圧縮率がEy成分に比べて悪くなっている．

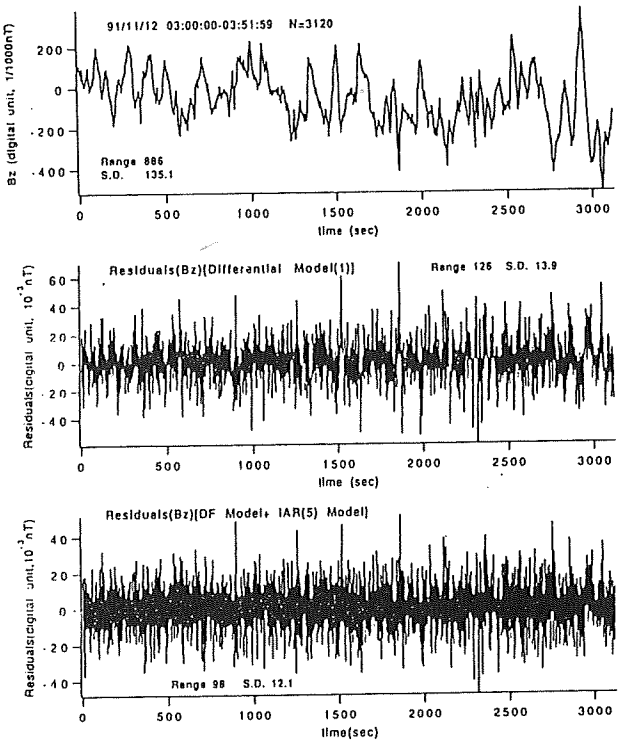

図3．磁場データと残差データ

以上のように生データに直接に幾種類かの圧縮法を適用しても圧縮率はそれほど望めないことがわかる．そこで，データにある種のモデルを当てはめ，データの取り得る値の範囲を小さくすることを考える．ここでは，以下に示す線形予測モデル（差分モデルと整数化ARモデル）で長周期変動を吸収し，取りえる値の範囲の小さな時系列に変換して各種の圧縮方法で圧縮をすることを考える．

1）差分モデル（DFモデル）

$$e_t = x_t - x_{t-1}$$

差分を繰返して残差データ中の0の数が一番おおくなったところで処理を終了するというアルゴリズムを採用した[6]．ほとんどのデータで1階差分ないしは，2階差分で十分であった．

2）整数化ARモデル（IARモデル）

$$e_t = x_t - \mathrm{INT}(a_1 x_{t-1} + \ldots + a_5 x_{t-5})$$

a_iはAR係数である．ARモデルの次数は，本来AICなどを使って最適次数を決めるのがよいが，次数が大きくなると計算時間が飛躍的にふえるので，今回は一律に5次のARモデ

ルを当てはめることにする．ただし，生データに直接整数化ARモデルを当てはめるよりも，差分を１回ほどこしたデータに当てはめた方が残差が小さくなることがわかったので，ここでは１階差分のデータに５次のARモデルを当てはめている．また，メモリの使用効率を考え残差を整数化する変換をほどこしている．

各データに２つのモデルをあてはめた場合の，残差がどうなるかを図１～図５に示す．また，各々のデータには，残差の範囲と標準偏差の値を示してある．磁場データはモデルの当てはめにより，範囲・標準偏差のどちらも生データの１０分の１程度になっている．しかし，電場データについては分散はかなり小さくなっているものの範囲はわずかに小さくなっているだけである．これは，磁場データに比べ電場データにはランダム・ノイズ的な要素が含まれているためである．

各残差に各種の符号化を適用した結果を図６に示す．磁場データに関しては，残差の範囲が-128～127の範囲なので２バイト・データとしてではなく，１バイト・データとして取り扱っている．それで，磁場データで圧縮率が50％を越えているものは，圧縮に伴うヘッダー・データにより元のデータ量よりも大きくなっていることを示している．Lempel-Ziv符号化の圧縮率が悪いのは，線型モデルを当てはめたことにより残差の性質がラ

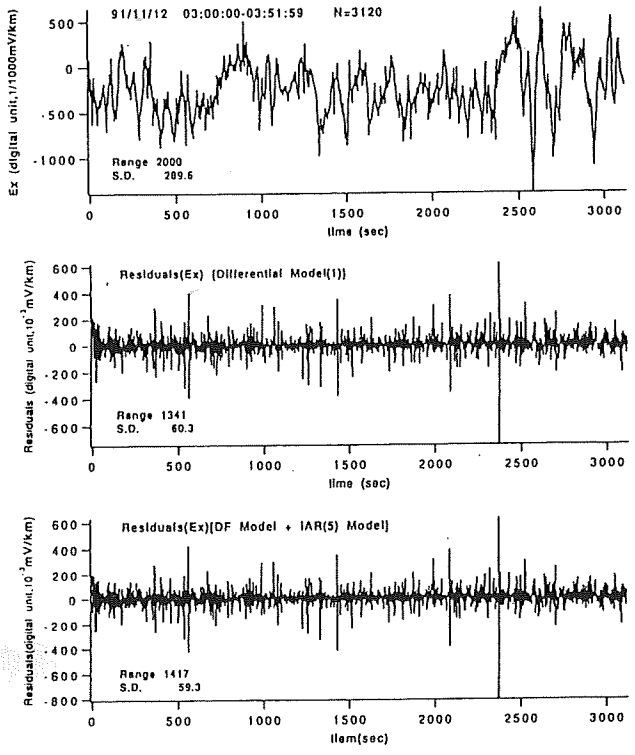

図４．電場データと残差データ

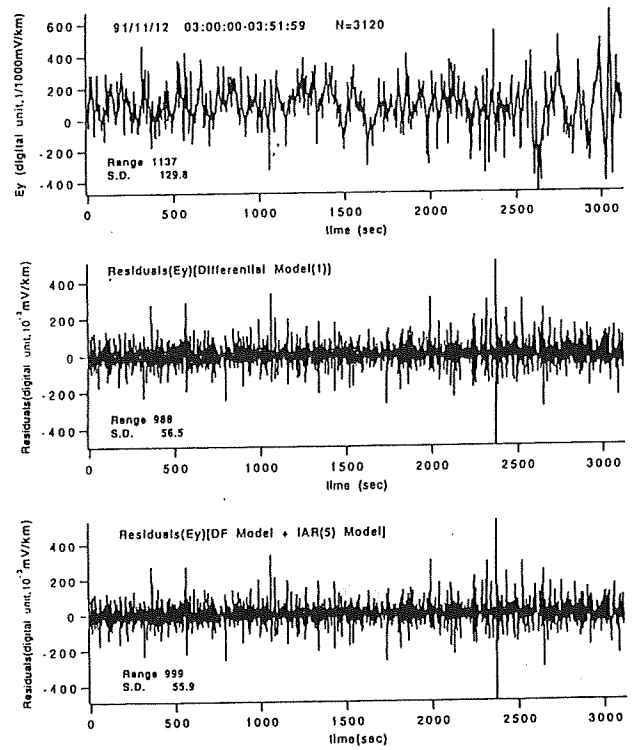

図５．電場データと残差データ

ンダムになったためと考えられる．圧縮率に関して，差分モデルと整数化ＡＲモデルとの違いはほとんどみられない．

一方，電場データの残差は範囲が広いために２バイト・データとして扱い，各種の符号化を適用した．差分モデルの当てはめによる残差データは，生データを圧縮する場合に比べてわずかに圧縮されている．しかし，整数化ARモデルを当てはめた残差の場合には，生データを圧縮した場合と変わらないか，かえって悪くなっている場合もみられる．また，符号化でいえば，磁場データと同じくLemple-Ziv符号化による圧縮率は悪くなっている．

今回検討に使用した磁場データに関しては、線型モデルを当てはめることによりデータを40％程度に圧縮できることがわかったが、電場データに関しては生データを圧縮した場合とさほど変化は見られなかった。差分モデルを当てはめた残差をLHaで圧縮した場合に数％の改善が見られたのみである．これは、もともと電場データにランダム・ノイズ的な要素が多いために今回適用した線形モデルでは十分にその特性を表現できていないためと思われる．

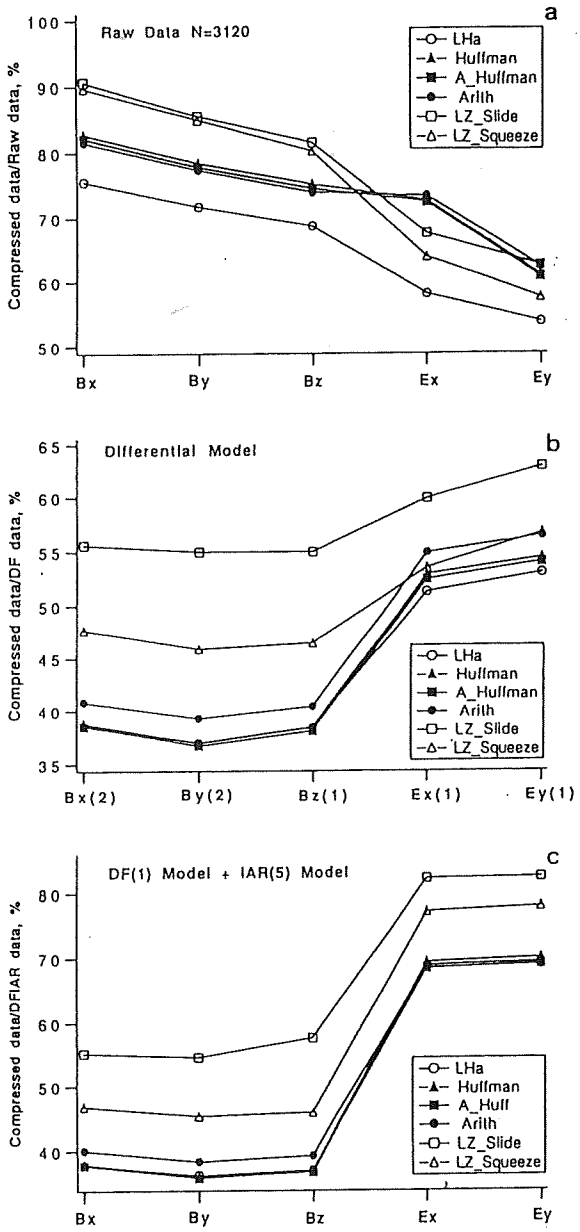

図６．符号化による圧縮率
a：生データ，b：差分モデル，
c：整数化ARモデル

４．まとめ

MT法で取得される磁場３成分と電場２成分について，線形モデルの当てはめをおこないその残差を圧縮するということを試みた．今回検討した範囲内では，簡単な差分を取るだけ十分であり，それ以上の処理をしてもそれほどの改善はみられなかった．また，符号化については，磁場データと電場データの性質によりかなりことなった結果となった．もともとランダム・ノイズ的な要素が強く変動幅の大きな電場データの場合は，モデルを適

用するよりも適当な符号化（Lempel-Ziv符号化など）を直接適用した方が良い場合があることがわかった．磁場データについては，長周期変動が卓越しているので線形モデルを使い時系列データを範囲の小さな残差時系列に変換してやることでかなりの圧縮ができることがわかった．この場合には，統計的符号化（Huffman符号化など）が有効であることもわかった．

今回評価に使った磁場データの変動は，比較的変動の範囲が狭く線形モデルの当てはめにより残差が-127～128の間におさまったので圧縮率が非常に高くなったが，電場データのように変動が大きい場合には２バイト・データとして扱う必要があるので圧縮率はかなり落ちると考えられる．ただし，この場合でもすべてのデータが２バイトを必要とするわけではないので，データ表現の無駄な部分をさらに圧縮する方法の開発の必要がある．

また，観測機器に圧縮機能を持たせる場合には，一括処理をするのかリアルタイムで圧縮をするのかという問題がある．例えば，今回使ったデータ（6240バイト）に差分モデルを当てはめ，残差をHuffman符号化するという処理を32ビットCPU80386SX（12MHz）で処理した場合，約１秒かかる．6240バイトというのは，データ数にして3120個である．つまり，１Hzサンプリングで24時間分の１成分のデータを処理するのに約28秒かかることになる．実際のサンプリング・レートが１Hzよりも遅いという場合には，CPUの稼働時間にかなりの余裕があるはずなのでリアルタイムの圧縮処理も可能と考えられる．この場合には，リアルタイムで差分をおこない動的符号化（動的Huffman符号化，Lempel-Ziv符号化）をおこなうという方法が考えられる．また，ネットワークでの配布等を考えた場合には，処理時間よりも圧縮率が問題になる．この場合には，磁場３成分あるいは電場２成分の相互相関を利用した処理も考えられ，これは今後の課題と思われる．

参考文献

1)Wood,L., Seismic data compression methods, Geophysics, 39, 499-525,1974.

2)Murakami,H. and Mizutani,H., Preliminary report:Data compression for LUNAR-A Penetrator, Proc. 25th ISAS Lunar Planet. Symp., 238-242, 1993.

3)地殻比抵抗研究グループ，滋賀県北西部に置ける地球電磁気共同観測，地磁気観測所技術報告 特別号（ＣＡシンポジューム講演論文集），32，71-75，1992.

4)Nelson, M., The data compression book, M&T Books, 1991.

5)奥村晴彦，Ｃ言語による最新アルゴリズム辞典，技術評論社，1991.

6)鷹野　登・武尾　実・高橋正義・三浦勝美・坪井誠司・阿部勝征，Fibonacci波形圧縮法を用いた地震波データの連続収録，地震学会予稿集　1990年秋季大会，p.280.