地球電磁気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 Geomagnetic and Geoelectric Data

MURAKAMI Hideki

Department of Geology, Faculty of Science, Kochi University

1．はじめに

地球電磁気学における全磁力，地磁気３成分，電場データといった観測データの量は，観測点の増加，測定方法や測定機器の改良に伴い非常に増えている．近年では，これらのデータの交換をインターネット等の計算機ネットーワークを介しておこなう機会が多くなってきている．しかし現状では，ネットワークへの負荷，あるいはデータ・サーバーとなる計算機の記憶媒体の容量の確保といった問題から生データをそのままの形式で記憶・転送することには問題がある．一般的に，大量のデータは計算機上でcompress，LHaなどに代表される汎用圧縮ソフトで圧縮された形式で保存されていることが多い．しかし，データの性質を積極的に利用すれば，これらの汎用圧縮ソフトで得られる圧縮率よりもさらに高い圧縮率を得ることが可能である．また，観測機器の機能としてリアルタイムでの圧縮機能があれば，現状のハードウェア規模で観測期間が延ばせるなどのメリットが考えられる．特に，火山地域の臨時観測や海底観測などにおいては観測計画等を立てる際の自由度がもたせられるなどのメリットがある．

以上のような点から，ＭＴ法で取得される地磁気３成分と電場２成分のデータの圧縮方法について検討した結果を前回報告した[1]．今回は前回の報告の中で問題になった点についてプログラムの改良をおこない検討をしたので，その結果について報告する．また，地磁気全磁力のデータについても検討をしたのであわせて報告する．

2．データ符号化プログラムの改善

前回の報告では，ＭＴ法で取得する磁場３成分と電場２成分のデータにたいして各種の圧縮方法を適用し，比較をおこなった結果次のようなことがわかっている[1]：１）磁場データは，サンプリングレートに対して比較的低周波成分のパワーが大きいので，生データを圧縮するよりも線型モデル（差分モデルあるいは整数化ARモデル）をあてはめた後の残差を符号化するほうが高い圧縮率が得られる；２）電場データの場合には，ランダムなノイズ成分を含む場合が多く，その場合には線型モデルをあてはめた残差と生データを直接に符号化した場合との差は比較的小さく，整数化ARモデルの場合にはかえって悪くなるケースもあった；３）符号化については，Huffman符号化，算術符号化，動的Huffman符号化，Lempel-Ziv符号化等について検討したが，モデルを当てはめた残差はランダムな時系列になることから，出現確率をもとに符号化するHuffman符号化が一般的に高圧縮率を与えることがわかった．今回は符号化に関してはHuffman符号化のみを検討する．

前回の検討では，一般的にパソコン等で使われているデータの入力が8bitsに限定されたプログラムを使用したために，次のような問題点があった：生データが16bitsであっても線型モデルをあてはめると残差のほとんどが8bitsであらわされる範囲にはいるが，すべてのデータがそうならないので残差

データも生データとおなじく16bitsデータとして扱う必要があり，圧縮という観点からみると無駄な部分が生じる．つまり，残差データすべてが8bitsで表現できる範囲におさまれば特に圧縮という処理をしなくてもそれだけでデータ量は半分になるが，全ての残差が8bitsの範囲におさまらない場合にはすべての残差データを16bitsデータとして処理する必要がある．そのために，本来はビット数合わせのための部分も情報と同じだけのウェイトを持って処理される．今回は，Huffman符号化プログラムについてこの点を改善するために，データの頻度分布を求める部分を8bitsから12bitsに拡張し，残差あるいは生データの変動分の範囲が-1024から1023までのデータを扱えるように改良した．12bitsに拡張したのは，使用しているパソコンならびにコンパイラのメモリー制限のためである．16bitsへの拡張は原理的には困難なことではないが，作業用のメモリを約1.6Mbytes確保する必要があり実用的ではない[2]．12bitsの場合には98 kbytes程度のメモリが確保できればよいので，計測機器などに使用されている16bits CPUなんどでも現実的と考えられる．

3．結果

評価には，前回の報告で使用したのと同じく1991年に実施された琵琶湖西部の総合観測で三国峠においてＵ３０によって得られたＭＴ法データを使用した．Ｕ３０では，１サンプルが16bitsデータとして取得されている．

今回使用したデータでは，データの変動分の範囲が12bits以下であったので，まず生データを直接にHuffman符号化をしてみた．その結果を第１表に示す．第１表には，比較のために8bitsのHuffman符号化プログラムと汎用の圧縮ソフトウェアLHaを適用した時の結果も示す．

第１表　生データの圧縮

		12bits Huffman		8bits Huffman		LHa
電場	Ex	3248 bytes	(472)	4541 bytes	(186)	3634 bytes
	Ey	2718	(304)	3815	(172)	3360
磁場	Bx	6737	(2821)	5156	(323)	4720
	By	5701	(2000)	4899	(323)	4482
	Bz	4864	(1381)	4701	(323)	4293

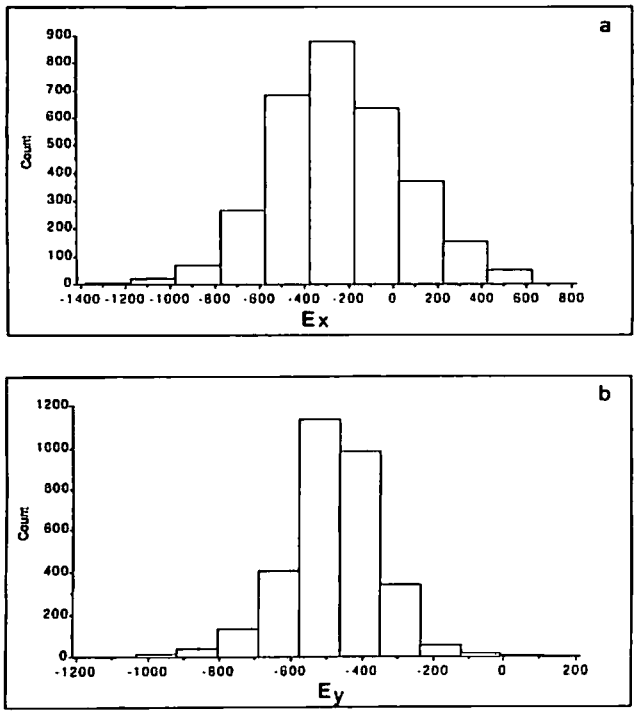
元データのサイズはぞれぞれ6240bytesである．括弧内の数字は
ヘッダー情報のサイズを表わしている．

まず8bitsHuffman符号化プログラムと12bitsHuffman符号化プログラムの結果を比較すると，電場データに関しては圧縮率が20％程度よくなっているが，磁場データについては圧縮率が逆に悪くなっていることがわかる．磁場のＸ成分（Bx）は生データよりもファイルサイズが大きくなっている．この原因は12bitsデータが取り扱えるようにしたために解凍時に必要なヘッダー情報が大きくなったためである．第１表の8bitsHuffmanおよび12bitsHuffmanの結果の欄の括弧の中の数字は各々のヘッダー情報のサイズを表している．ヘッダーサイズを除いたデータ部の実質のサイズは8bitsHuffmanのときよりもかなり小さくなっていることがわかる．次に，汎用圧縮プログラムLHaとの比較でみると，圧縮ファイルのサイズだけで比較するとＬHaの方が小さいくなっている．LHaのヘッダーサイズがどの程度か

は分からないが，8bitsHuffmanのヘッダーサイズ程度とすると，ヘッダー情報を除いた部分のサイズは12bitsHuffman符号化プログラムの結果の方が小さくなっている可能性がある．

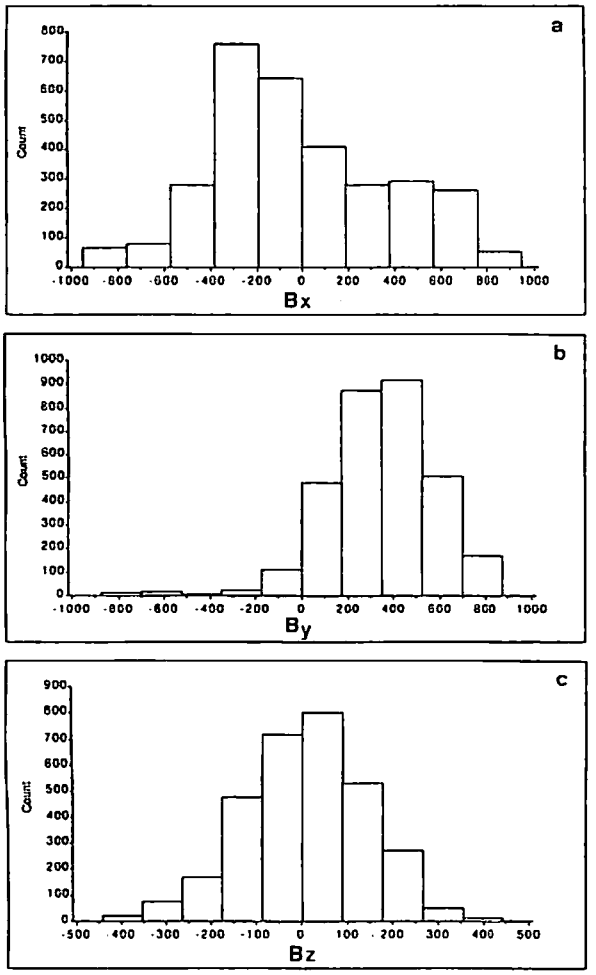
圧縮プログラムとしての完成度を高める意味では，ヘッダー情報の圧縮について検討する必要がある．汎用圧縮ソフトウェアLHaでは，Lempel-Ziv(sliding dictionary)法とHuffman法の組み合せというだけでなく，圧縮率を向上させるためにヘッダー情報をも圧縮している[3]．

生データを12bitsHuffman符号化プログラムで圧縮した場合でも，電場データは8bitsHuffman符号化のときよりも圧縮されているのになぜ磁場データは大きくなったのかを検討するために，扱った生データの振幅の頻度分布図を作成してみた．その頻度分布図を第１図と第２図に示す．第１図は電場２成分の振幅の頻度分布を示しているが，Ex，Eyともに比較的きれいなガウス型の分布をしている．第２図には磁場３成分の振幅の頻度分布を示す．一見してわかるように左右が対称なガウス型の分布をしていない．このようなデータ構造をしているためにHuffman treeのバランスが悪くなり，ヘッダー情報が大きくなっていると考えられる．その証拠に磁場３成分のなかではガウス型の分布をしているBz成分のヘッダー情報が他の成分に比べて小さくなっている．



第１図　電場データの振幅ヒストグラム

次に差分モデルを当てはめた残差について，12bitsHuffman符号化プログラムを適用した結果を第２表に示す．第２表には比較のために8bitsHuffmanを適用した結果も示す．ただし，8bitsHuffmanの結果のうち磁場データは残差が8bitsの範囲におさまっているので，圧縮前の残差データのサイズがすでに半分になっている状態で符号化をおこなっている．



第２図　磁場データの振幅ヒストグラム

		12bit Huffman	8bits Huffman
電場	Ex(1)	2355 bytes	3302 bytes
	Ey(1)	2450	3394
磁場	Bx(2)	2518	2420
	By(2)	2390	2306
	Bz(1)	2505	2403

第２表　差分モデルあてはめによる残差データの圧縮

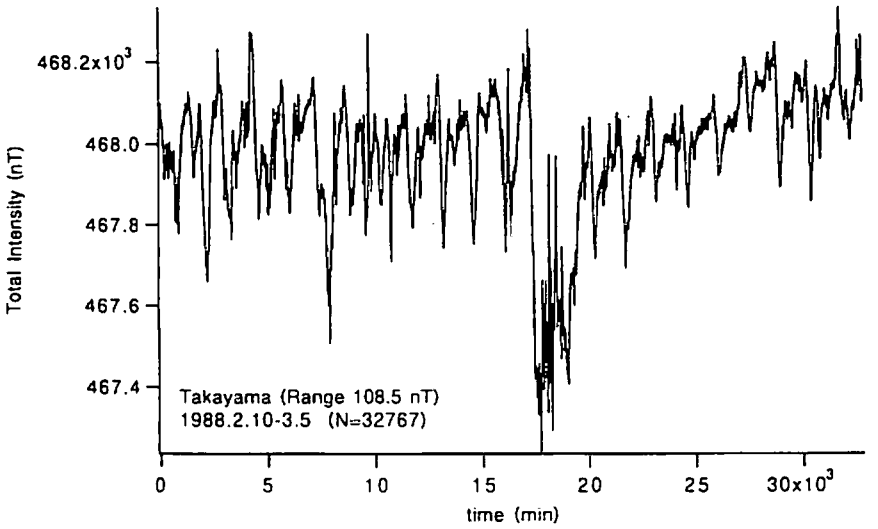
成分表示の隣の括弧の中の数字は差分の回数を表わしている．

また，磁場成分の8bits Huffmanの結果は残差が8bits以内であったので

１データが8bitsとして扱っている．

第２表の結果をみると電場については12bitsHuffman符号化プログラムの結果の方が良く，磁場については8bitsHuffmanの結果の方が良くなっているがこれは上に述べたような理由により小さくなっている．しかし，その差は生データの場合の差と比べると大変小さくなっており，差分モデルを適用した効果が出ている．生データでは非対称な分布をしていた振幅分布が，残差では対称なガウス型の分布になるのでHuffman treeのバランスがよくなりヘッダー情報が小さくなり，同時にデータの分散が小さくなることにより出力されるコードサイズ自体も小さくなっている．その結果，電場データも磁場データも生データの４０％程度にまで圧縮されている．

次に地磁気全磁力データについて検討をおこなった．使用したデータは，名古屋大学高山地震観測所において観測されている1988年2月10日から3月5日までの地磁気全磁力の毎分値データ32767個である．全磁力データの変化の様子を第３図に示す．データの範囲は108.5nTで比較的静穏な期間のデータといえる．



第３図　高山地震観測所における全磁力データ

１階差分+Huffman Coding	15759 bytes
整数化ＡＲモデル+Huffman Coding	
３次ＡＲモデル	15588
５次ＡＲモデル	15312
１０次 ＡＲモデル	15282

第３表　全磁力データの各モデルによる圧縮

元のデータ量は 64kbytes である．

このデータの変化分だけを2bytes(16bits)データとして記憶すると，ちょうど64kbytes必要になる．前回ＭＴ法データにおこなったように線型モデルをあてはめ，その残差を符号化するという手順で評価をおこなった．全磁力データの毎分値の場合には，変動がそれほど大きくないので残差がどのモデルの場合でも8bitsで表現できる範囲内（-12.8nT～12.7nT）におさまるので8bitsHuffmanを適用した．その結果を第３表に示すが，いずれのモデルの場合でも生データの約２５％程度になっている．また，差分モデルと整数化ＡＲモデルを比較すると，整数化ＡＲモデルの方が圧縮率は高く適用するＡＲモデルの次数を高くすればさらに圧縮率が高くなることがわかる．しかし，その差は小さくＡＲモデルの次数を高くすると計算時間が飛躍的にかかることを考慮すると一番単純な１階差分モデルが現実的と考えられる．

４．まとめ

今回は，Huffman 符号化プログラムを扱うデータのダイナミックレンジに対応するように拡張して12bits データまで直接扱えるようにした．MT法データについて検討した結果，生データに適用する場合にはデータの構造によりヘッダー情報が大きくなり圧縮にならないケースもあるが，線形モデル（差分モデルや整数化ＡＲモデルなど）を適用してランダムな時系列に変化してやれば拡張したHuffman Coding プログラムは有効に作用することがわかった．評価に使用したMT法データでは，電場・磁場データのいずれも４０％程度まで圧縮できた．しかし，ヘッダー情報の圧縮について今後検討すれば圧縮率という面ではさらに改善される可能性がある．

また，全磁力データのようにサンプリングレート（毎分）に対して比較的緩やかに変化するようなデータの場合には，１階差分とHuffman 符号化という組み合わせでも約２５％まで圧縮可能であることがわかった．これは，固定観測点におけるデータの保管には役立つものと思われる．

これまでは，Huffman 符号化による圧縮率を高くするために，生データに簡単な線形モデルをあてはめて取り扱うデータの範囲と分散を小さくするという方法を取ってきた．しかし，これまでに扱ってきた線形モデルがいつもうまくいくわけではなく，データの変動にたいして十分なサンプリングレートでない場合には，単純差分を適用した場合には範囲が逆に広くなる場合があり生データを直接圧縮した方が良い場合もある．このような問題があるので，電場・磁場についての特性をうまく抽出するモデルについてはさらに検討する必要がある．

参考文献

１）村上英記，MT法データの圧縮について，1993年ＣＡ研究会論文集，1-6，1993．
２）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．
３）奥村晴彦・吉崎栄泰，圧縮アルゴリズム入門，C Magazine，44-68，1991．