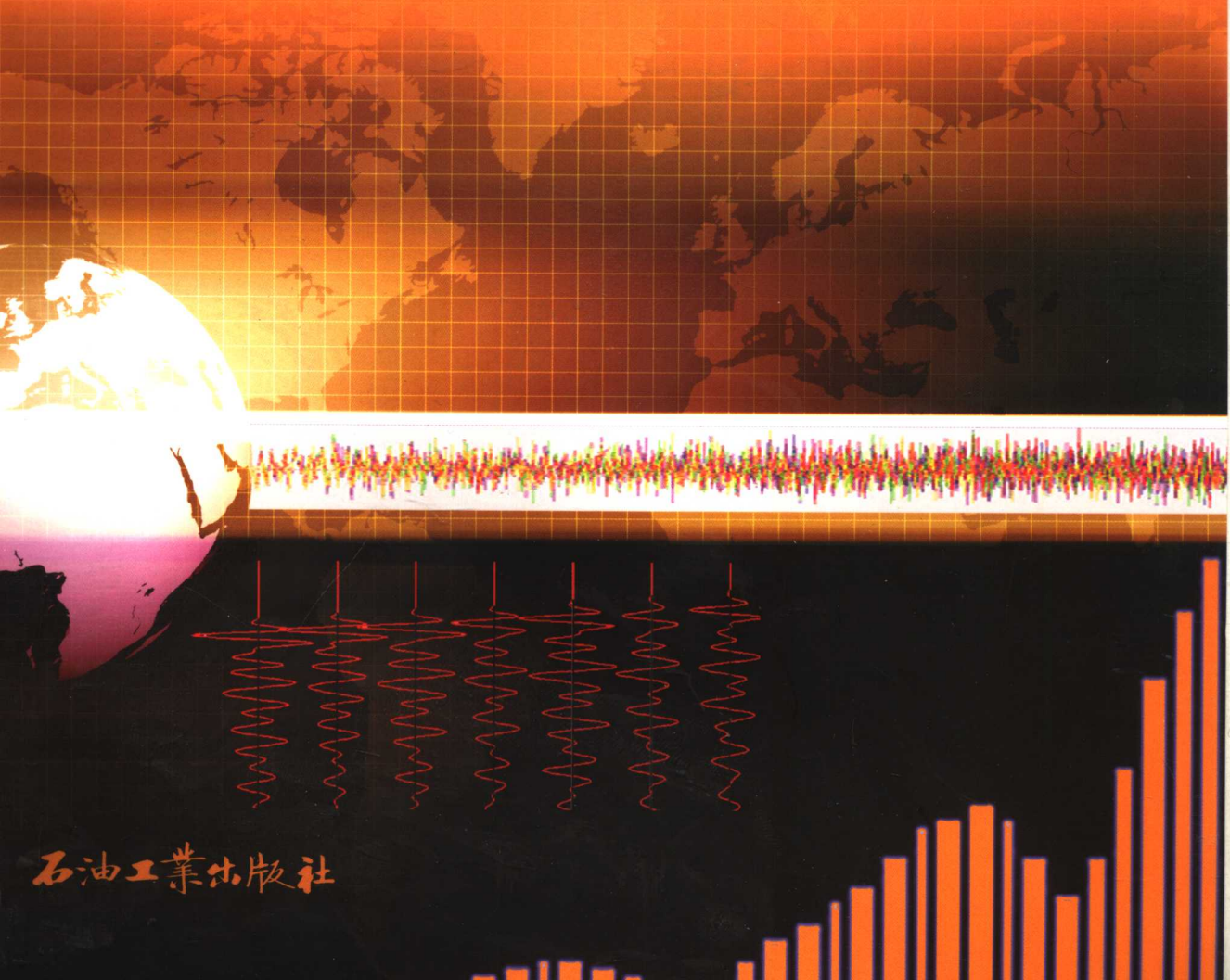




DIZHEN SHUJU YASUO FANGFA YU YINGYONG

地震数据压缩 方法与应用

武文波 胡福祥 徐 凌 秦前清 编著



石油工业出版社

地震数据压缩方法与应用

武文波 胡福祥 徐 凌 秦前清 编著

石油工业出版社

图书在版编目 (CIP) 数据

地震数据压缩方法与应用/武文波等编著.

北京:石油工业出版社,2007.8

ISBN 978-7-5021-6220-7

I. 地…

II. 武…

III. 地震数据-数据压缩

IV. P315.61

中国版本图书馆 CIP 数据核字 (2007) 第 123172 号

出版发行:石油工业出版社

(北京安定门外安华里 2 区 1 号 100011)

网 址:www.petropub.cn

发行部:(010) 64210392

经 销:全国新华书店

印 刷:北京晨旭印刷厂

2007 年 8 月第 1 版 2007 年 8 月第 1 次印刷

787×1092 毫米 开本:1/16 印张:6.75

字数:176 千字 印数:1—1000 册

定价:28.00 元

(如出现印装质量问题,我社发行部负责调换)

版权所有,翻印必究

前 言

地震勘探是人类获取地球内部信息的重要手段，是寻找地下油气和其他矿物的基本方法之一。石油是不可再生的资源。近几年来，随着世界各国经济的发展，石油短缺问题已成为影响各国经济安全和可持续发展的首要问题。面对严峻的国际形势和世界的发展趋势，国家在“十一五”科技发展计划中明确提出：重点研究西部复杂地质条件下油气、煤层气和深海油气资源的高精度地震勘探和开采技术。当前，地震勘探技术正向多维、多分量、多参数、高分辨率方向发展，使得勘探数据呈指数趋势逐年增加，地震数据量不断膨胀。然而，在获取这些海量观测数据之后如何对其进行有效传送和存取是当前制约这些数据被进一步开发、利用的重大难题。按照常规的方法处理，这些数据从现场采集到运回处理中心做分析解译往往需要几个月的时间。如果分析结果反映出地震数据采集有问题，就需要重新激发放炮，这将会造成大量时间和物资的损失。传输手段的滞后严重影响到油气勘探的效率。

海量数据压缩作为信息获取与处理的一项关键技术，在过去的几年内得到了良好的发展。不过，总体而言，数据压缩技术在地震勘探中的应用尚处于初级阶段。随着计算机技术和勘探技术的发展，地震数据压缩必将得到进一步的发展和應用。数据压缩是解决海量数据传输和存储的一项关键技术。通过先进的压缩算法可以大大降低传输的时间，提高数据传输的可靠性，降低传输成本，同时基本不影响勘探数据的质量。因此，对海量地震数据进行有效压缩是当前亟待解决的技术关键。

针对地震勘探技术的发展现状，本书以地震勘探获取的实际数据为研究对象，介绍了海量地震数据压缩的前沿技术。第一章是地震数据压缩编码技术的概述，重点阐述数据压缩的必要性和可行性，以及地震数据压缩编码的内容。第二章对地震数据压缩的理论基础——地震信号的特点、变换理论（离散余弦变换、重叠变换、小波变换等）作了简洁而重点的介绍。第三章详细讨论了地震数据无损压缩的编码方案，包括基于预测方法和基于整数小波变换两类浮点地震数据的无损压缩方法，并给出了具体的实现过程。第四章主要介绍原始地震数据的高保真无损压缩问题，针对地震数据的特点，基于重叠变换和小波变换技术，给出了具体的编码方案。第五章对压缩数据的质量进行评估，并对地震数据压缩提出参考意见。在本书中，笔者将根据自己从事地震数据压缩编码研究的成果与体会，在这方面做出努力，旨在帮助读者透彻理解和掌握地震信号压缩的基本原理和框架，提高正确选择、使用、设计实现信号压缩编码方案和系统的能力，了解本学科领域的发展趋势和有待进一步深入研究的问题。

本书参考和引用了很多学者的研究成果和论著，笔者谨对这些文献资料的著者表示诚挚的谢意。鉴于本学科涉及知识面非常宽广，限于笔者的学术水平，不妥之处在所难免，敬请读者不吝赐教。

目 录

1 绪论	(1)
1.1 地震数据压缩	(1)
1.2 地震数据压缩的信息论基础	(3)
1.3 地震数据压缩的原理	(4)
1.3.1 编码冗余	(4)
1.3.2 空间冗余	(5)
1.3.3 心理视觉冗余	(5)
2 地震信号压缩中的变换方法	(7)
2.1 概述	(7)
2.2 地震信号分析	(8)
2.2.1 地震信号的信息量	(8)
2.2.2 地震信号的相关性分析	(9)
2.2.3 地震信号的功率谱分析	(10)
2.2.4 地震信号的模型分析	(11)
2.3 余弦变换	(15)
2.3.1 一维 DCT 变换	(15)
2.3.2 二维 DCT 变换	(16)
2.3.3 重叠正交变换 (Lapped Orthogonal Transform, LOT)	(18)
2.3.4 LOT 的优化及快速计算	(19)
2.4 小波变换	(21)
2.4.1 Gabor 变换及其局限性	(22)
2.4.2 连续小波变换与离散小波变换	(22)
2.4.3 多分辨率分析	(25)
2.4.4 小波变换的快速算法	(26)
2.4.5 信号的小波分解与重构	(28)
3 地震数据无损压缩	(32)
3.1 概述	(32)
3.1.1 HUFFMAN 编码	(32)
3.1.2 算术编码	(34)
3.1.3 字典编码	(37)
3.1.4 游程编码	(37)
3.2 基于预测编码的浮点地震数据无损压缩	(38)
3.2.1 几种常用的预测器	(39)
3.2.2 自适应方向预测器	(41)
3.2.3 压缩编码实现	(42)

3.2.4	实验分析	(44)
3.3	基于整数小波变换的浮点地震数据无损压缩	(46)
3.3.1	整数小波变换	(46)
3.3.2	整数小波变换的快速实现	(51)
3.3.3	数据预处理	(52)
3.3.4	压缩编码实现	(53)
3.3.5	实验分析	(54)
4	地震数据有损压缩	(55)
4.1	概述	(55)
4.2	地震数据的预处理	(56)
4.2.1	地震信号特征	(56)
4.2.2	地震数据分块处理	(56)
4.2.3	地震原始数据增益调整	(57)
4.3	量化	(59)
4.3.1	标量量化	(59)
4.3.2	矢量量化	(59)
4.3.3	地震信号压缩的量化设计	(60)
4.4	基于重叠变换的地震数据有损压缩	(62)
4.4.1	多分辨率重叠双正交变换 (Hierarchical Lapped Biorthogonal Transform)	(62)
4.4.2	量化和系数组织	(63)
4.4.3	熵编码	(64)
4.4.4	实验分析	(64)
4.5	基于小波的地震数据有损压缩	(68)
4.5.1	小波基的选取	(68)
4.5.2	小波分解的层数	(70)
4.5.3	基于线的小波变换算法 (Line-based Wavelet Transform, LWT)	(70)
4.5.4	位平面编码	(70)
4.5.5	小波系数的熵编码	(73)
4.5.6	实验分析	(74)
5	地震数据压缩质量评价	(79)
5.1	概述	(79)
5.2	地震数据质量评价	(80)
5.2.1	主观评价信息的定义和度量	(80)
5.2.2	客观评价方法	(80)
5.3	压缩对地震数据处理的影响	(91)
5.3.1	地震数据处理过程	(92)
5.3.2	压缩对处理影响的主观评价	(93)
5.3.3	压缩对处理影响的客观评价	(94)
	参考文献	(100)

1 绪 论

20 世纪是一个令人目不暇接的世纪，其间上演的一幕幕科学和技术的神话，以前所未有的方式改变着人类的生活。迈入 21 世纪，我们有理由相信科学技术作为人类社会发展的巨大推动力必将加速改变我们的生产和生活。自远古人类结绳记事到 1901 年意大利发明家马可尼成功地试验了跨越大西洋的无线电报，再到 1946 年第一台电子计算机在宾夕法尼亚大学摩尔学院问世，之后经历了大型机时代，PC 机时代，过渡到互联网时代。伴随着网络通信和计算机技术的全面发展，人类社会从此进入了信息时代。

1.1 地震数据压缩

信息时代的一个重要特征是信息大爆炸。信息的膨胀集中体现在表达信息的数据量的增长。在过去的十几年里，数据传输和存储技术的进步不啻于一日千里，然而海量数据的快速增长经常使技术的进步销弥殆尽。现在，一些高性能工作站的内存都用 GB 来衡量。外存储设备的发展同样惊人，低速率低密度的纸带打孔机先被淘汰，后来 5.25in, 360KB 的低密软盘又过时，取而代之的是 10MB 的优盘，发展到现在，普通的 PC 机硬盘的容量都是 100GB。在存储设备发展的同时，现代通信技术也有了长足的进步，从模拟到数字，从不足 1Kbps 的点对点通信到主干网带宽达到 1Tbps 的光纤通信。短短十几年，技术的变革如此巨大，这在人类历史上是罕见的。可是即便如此，那些码放得整整齐齐，黑压压堆满整座大楼的高密度磁带仍然时常让人心悸。问题在于，通信和存储技术的发展总也赶不上数据的增长速率。

石油是现代工业的血液，在国民经济建设中具有无比重要的战略作用。地震勘探是人类获取地球内部信息的重要手段，是寻找地下油气和其他矿物的基本方法之一。地震勘探是石油勘探方法中最为重要和目前应用最多的一种地球物理勘探方法。它是通过人工方法激发地震波，研究地震波在地层中传播的情况，以查明地下的地质构造，为寻找油气田或为其他勘探目的服务的一种物探方法。自 1927 年地震勘探成为一种常规勘探方法以来，地震勘探技术始终处于不断发展进步之中。先后经历了光点地震仪勘探阶段、模拟磁带仪勘探阶段、数字地震仪勘探阶段，目前正向着高分辨率、高精度方向发展。在所有石油勘探数据中地震勘探数据量是最大的。由于地震勘探实现数字化最早、数字化程度最高，因而其数据保存在世界范围内具有共同的性质，而且随着数据量的逐年积累，旧有的数据管理与保存模式已不能适应现代生产的需求，在数据存储、管理、数据检索与使用方面产生了许多问题。

海量数据的传输和存储是信息时代普遍存在的重大难题。在石油部门，随着石油勘探工作的不断深入，高分辨率三维地震等先进技术的广泛应用，地震勘探正向多维、多分量、多参数、高分辨率方向发展，使得勘探数据呈指数趋势逐年增加，地震数据量不断膨胀。然而，在获取这些海量观测数据之后如何对其进行有效传送和存取是当前制约这些数据被进一步开发、利用的重大难题。按照常规的方法处理，这些数据从现场采集到运回处理中心做分析解译往往需要几个月的时间。传输手段的滞后严重影响到油气勘探的效率，如果分析结果

反映出地震数据采集有问题，就需要重新激发放炮，这将会造成大量时间和物资的损失。有关专家做过估算，在全球地震勘探中，这种决策延缓造成的浪费平均每月达 10 万美元。因而石油公司也开始采用通讯网络传输，如在海洋地震勘探和边远地区的地震勘探中，地震勘探数据要通过卫星或网络传输到基地进行分析处理。然而不断增长的数据量迫使我们不得不在现场对采集的数据压缩后再传送。另外，不断增长的数据量对地震资料的存储也构成了挑战。例如，在挪威水域进行一次典型的三维地震测量大概要覆盖 $1.5 \times 10^4 \text{ km}^2$ 的面积，产生约 1 万盘磁带、2T 字节左右的数据。过去地震数据都是存储在磁带上，由于时间过长、保存条件和磁带质量等原因，部分磁带出现粘连，造成了大量数据损坏或丢失。如何安全、有效地存储这些数据是当前亟待解决的问题之一。

这样，为了降低存储和传输的成本，开发有效的地震数据压缩算法成为一项亟待解决的技术。如果能将地震数据压缩十倍，那么存储空间将会节约 90%，传输时的带宽也会变成原来的十分之一，这将是一件非常吸引人的工作。很长时间以来，有人一直对地震数据压缩技术持有怀疑态度，在他们看来，庞大的地震数据量问题完全可以通过宽带技术和存储优化技术得以解决。的确，当今飞速发展的宽带技术和存储技术削弱了人们对数据压缩技术的需求，但是，不可否认，压缩技术目前来说仍然是非常重要的。一个简单的例子是可以从海上勘探船上实时地将勘探压缩数据传回基地处理中心进行实时处理 (Donoho 等, 1995; Stigant 等, 1995)。更普遍的是，我们利用的带宽或存储资源总是遵从帕金森定律：能够可用的资源总是满足不了人们对资源的需求。这样，尽管网络带宽和存储的花费呈现下降的趋势，然而从使用效率的角度来说，压缩技术的应用仍然有巨大的优势。

压缩技术可以分为两大类：无损压缩和有损压缩。所谓无损压缩是指在压缩和解压的过程中没有任何信息的丢失，可以从压缩数据完美重建原始数据。而有损压缩，意味着在压缩的过程中会造成部分信息的丢失。人们总是追求完美，因此，使用者更倾向于无损压缩。遗憾的是，无损压缩目前能够达到的压缩倍率大约在二倍左右，难以满足需求。对地震数据来说，在采样和记录的过程中，由于受到噪声的干扰，不可避免地引入了一些误差。因此，只要能够满足目前处理的要求，采用高倍率的有损压缩是完全可行的。

尽管理论上讲采用有损压缩是可行的，然而对地震数据进行有损压缩仍然没有成为一个普遍的共识。主要原因在于地震勘探目前正致力于如何提高数据的质量，而有损压缩潜在地会降低数据的质量。而近几年有研究指出 (Hewlett 和 Hatton, 1995; Hall 等, 1995)，对地震数据压缩 10 倍而引入的误差远远小于对地震数据进行其他常规处理造成的误差，如反卷积处理、叠加处理等。这些研究表明，有损压缩技术完全可以应用到处理和解释的阶段中去，尽管原始数据仍然需要完整保存。另外一个制约地震数据压缩技术实际应用的因素是压缩和解压的计算开销问题，然而随着计算机性能的飞速提高，这个问题基本上可以不用考虑了。这样，地震数据压缩技术可以在实际应用中找到很多的切入点。

地震数据准无损技术一旦应用于生产实践，将会带来巨大的经济和社会效益：

(1) 可以缩短资料处理的周期，带来间接的效益。在现有通讯条件下，为了使通讯频道或网络能够及时地传输高精度的地震数据，高保真压缩是一种实用的解决方法，该方法可以大大减少数据量，降低传输设备的投资成本，缩短传送特别是后续处理的时间，及时提供地震勘探成果，加快了整个勘探进度。经过压缩后的地震数据由于数据质量良好，消除了数据中的错误，数据与文档配套，减少了寻找和验证数据质量的时间，因此计算中心在重新处理时感到使用快捷方便，缩短了处理时间，节约了处理费用。

(2) 压缩后的数据存储体积大大减小, 减少了存储和保管费用。经过压缩后, 新数据重新保存在磁带或磁盘上, 存储量明显减少, 大大减少了磁带保存空间。原来用几千平方米的库房保存的数据现在可以用几十平方米就可以保存, 减少库房面积 90% 以上, 不但便利了资料管理, 也易于建立符合磁带保管的环境, 减少能源消耗和管理费用。如果用计算机进行存储, 可以为勘探数据的信息化、网络化奠定基础。西方勘探公司做过计算, 认为基于数据压缩技术使勘探成本呈线性下降趋势, 即 10 : 1 压缩比例成本是原来的 10%, 20 : 1 是原来的 5%。如果考虑到石油工业每年进行的 200~300 个区块的三维地震, 勘探成本将大大降低。

本选题就是在此背景下产生的。地震数据不同于一般的图像数据, 其数据压缩更有挑战性。与一般的图像数据相比, 地震数据的特性突出表现在以下几个方面 (王培茂, 2003): (1) 地震波形是紧支集的, 空间相关性和谱间相关性都很强, 函数的转换需要在空间域和频率域都有良好的局部性; (2) 地震信号分频带特性强, 在二维谱上常表现出很强的条带性。在水平方向和垂直方向频谱特性差异很大; (3) 地震信号的相位、幅度都包含着丰富的地层信息且动态范围大; (4) 压缩保真度要求高。地震信号处理最显著的特点之一, 就是要限制物理信息或信号特征的丢失, 特别是高频信息。因为需要根据这些信息判断地层结构、油气藏情况等。因此, 对海量地震数据进行有效压缩是当前亟待解决的技术关键。

1.2 地震数据压缩的信息论基础

直觉上我们知道什么是信息, 日常生活中我们通过文本、语言和图像会传递很多信息。但同时我们又觉得信息是个难以理解的不确定量, 无法精确定义、捕获和度量。信息论是应用近代数理统计方法来研究信息的传输和处理的科学。“信息论之父”美国工程师 C. E. Shannon 于 1948 发表了“通信的数学理论”, 这是一篇关于现代信息论的开创性权威论文, 该论文系统讨论了信源和信道特性, 给出了信息度量的数学公式, 为信息论的创立做出了独特的贡献。

信息论的重要作用在于它给出了信息的定义和度量方法。根据信息论的理论, 信息是事物发展的不确定性的变化。即人们对某一事物先后两次认识的差别, 或某人对某一事物获得了新的知识、新的消息。清华大学的常迥院士认为信息的来源有两方面: (1) 只有变化着的事物才有信息; (2) 只有尚未确定的事物才有信息。归纳而言: 信息是指对某一事物的新知识、新认识 (常迥, 2003)。

信息的本质是事物的不确定性的变化, 信息的定量表征必然联系着不确定性的度量。根据概率论, 事件的不确定程度, 可用其出现的概率来描述, 事件出现的可能性越小, 则概率就越小, 反之, 则概率就越大。消息中的信息量与消息发生的概率紧密相关, 消息出现的概率越小, 则消息中包含的信息量就越大。

令 S 代表一组事件 E_1, E_2, \dots, E_N , 它们出现的概率分别为 $P(E_i) = p_i, 0 \leq p_i \leq 1$, 并有

$$p_1 + p_2 + \dots + p_N = 1$$

事件 E_i 的自信息记作 $I(E_i)$, 并定义为

$$I(E_i) = -\log_2 p_i \quad (1.2.1)$$

p_i 愈小则 $I(E_i)$ 愈大, 这是和我们的感觉相符合的。某一事件越罕见, 则其出现所带来的信息就越多。

S 的熵, 称为 $H(S)$, 是自信息的统计平均值, 即

$$H(S) = E(I) = - \sum_{i=1}^N p_i \log_2 p_i \quad (1.2.2)$$

因为 p_i 可以为零, 使得定义中的 $p_i \log_2 p_i$ 成为不定式, 所以当 $p_i = 0$ 时, 指定 $p_i \log_2 p_i$ 等于零。一个事件的自信息是随其不确定程度的增长而加大的。所以熵可被认为是系统不确定程度的度量。当所有事件具有等概率时, 事件将在某一时间出现的不确定性为最大, 这时候 $H(S)$ 亦最大, 而对确定事件而言, $H(S)$ 为零。熵给出了数据在理论上可以达到的完全无冗余度压缩的极限值, 从而奠定了压缩理论的坚定基石 (张旭东, 2003)。

1.3 地震数据压缩的原理

一般说来, 自然界的数据或多或少都含有“水分”(即存在冗余), 数据压缩的目的就是把这些水分挤掉, 用最少的比特表示最多的信息。比如我们日常说的话, 70%以上都是多余的废话。通过一种运算, 把这些多余的东西剔除出去, 就实现了数据压缩。典型的数据压缩包含去相关(冗余)和熵编码两个步骤。实际上日常生活中拍发电报的过程就是一个数据压缩的经典范例。为了减少电报的字数, 我们把文字尽量精简, 浓缩为一种类似文言文的精练文字, 表达的意思保持不变; 然后将精简后的文字转换为莫尔斯电报码发送出去。这两个步骤分别相当于数据压缩的去相关和熵编码。

在数字信号(图像)处理中, 数据冗余可分为以下三类: 编码冗余、空间冗余和心理视觉冗余 (Rafael C. Gonzalez, Richard E. Woods, 2003)。采用不同的方法消除这些冗余, 可以使得数字信号(图像)得到有效的压缩。

1.3.1 编码冗余

编码冗余是一种和数据表示有关的冗余。对于数字信号来说, 在存储和传输时需要用一个码字来表示采样点, 如可以用 16 位或 32 位数来表示地震信号的幅度值。由于某些幅度值比其他幅度值有更大的出现可能性, 而在一般表示时对有最大可能性的值和最小可能性的值都分配相同的比特数。在这种情况下, 必然存在编码冗余。

假设区间 $[0, 1]$ 内的一个离散随机变量 r_k 表示信号的幅度值, 每个 r_k 出现的概率为 $p_r(r_k)$ 。

$$p_r(r_k) = \frac{n_k}{N}, \quad k = 0, 1, 2, \dots, L-1$$

这里 L 表示不同幅度值的个数, n_k 是第 k 个不同幅度值出现的次数, N 是信号的采样总数。如果表示每个 r_k 的比特数为 $l(r_k)$, 则表达每个采样点所需的平均比特数为

$$L_{avg} = \sum_{k=0}^{L-1} l(r_k) p_r(r_k)$$

由上式知表示该信号需要 $N \times L_{avg}$ 个比特的数据。通常在使用自然码直接编码时一般没有考虑数据值的概率分布特性, 故表示信号幅度值 r_k 的比特数 $l(r_k)$ 没有接近 L_{avg} , 从而产生

编码冗余。

例如在一个 8 灰度的图像中，灰度值及其概率分布如表 1.3.1 的第 1、第 2 列，如果分别采用 3 比特自然码和变长 Huffman 编码进行编码，码字分配如表 1.3.1 的第 3、第 5 列，则利用 3 比特自然码编码的 L_{avg} 为 3，利用 Huffman 编码的 L_{avg} 为 1.064。显然在利用自然码编码时比利用 Huffman 码编码包含有更多的编码冗余。

表 1.3.1 自然码及 Huffman 码编码对比

r_k	$p_r(r_k)$	自然码	码字长度	Huffman 码	码字长度
0	0.02	000	3	00111	5
1/7	0.06	001	3	00110	5
2/7	0.08	010	3	0010	4
3/7	0.12	011	3	011	3
4/7	0.14	100	3	010	3
5/7	0.16	101	3	000	3
6/7	0.20	110	3	11	2
1	0.22	111	3	10	2

1.3.2 空间冗余

对由于采样值出现概率不同引起的编码冗余，可以用变长编码等熵编码手段予以消除。然而，熵编码处理不能消除信号采样点之间相关性的级别，换句话说，用于表示采样值大小的编码与采样点之间的相关性无关，这些相关来自于信号中对象之间的结构或几何关系。

空间冗余是数字信号相邻采样点之间的统计相关性，有时也称为采样点冗余或几何冗余。对地震信号而言，由于地下地质构造和地质信息的连续性，使得获取的地震信号在相邻道或相邻采样点之间存在一定的相关性。信号的某个采样点可以在一定的误差范围内从周围的其他采样点预测出来。换句话说，单个采样点所携带的信息相对较少，它对信号整体的视觉贡献可能是多余的，这是因为它通常可以用其邻近的采样值来判断。

为了在编码时减少信号采样点之间的冗余（空间冗余），通常用于人类观察和解释的信号必须转换为更有效的形式，如采用预测变换、DCT 变换、小波变换等，通过这类变换可以消除采样点间的空间冗余。

1.3.3 心理视觉冗余

心理视觉冗余是不同于编码冗余和空间冗余的一种特殊的冗余，它跟人的感觉器官眼睛有关。眼睛作为感知外界世界的一个重要器官，不仅因人而异，而且所感受的区域亮度与区域和背景的反射光的强度都有关。如在灰度值为常数的区域也能感觉到灰度值的变化，这种现象的产生是由于眼睛并不是对所有视觉信息有相同的敏感度。有些信息在通常的视感觉过程中与另外一些信息相比来说不那么重要，这些信息可认为是心理视觉冗余的，去除这些信息并不会明显地降低所感受到的图像的质量。许多称之为“第二代编码技术”的有损方法就是基于这个原理的。另外，心理视觉冗余的存在与人观察图像的方式、思维等也有关系。人在观察图像时主要是寻找那些比较明显的感兴趣的特征，而不是定量地分析图像中每个像素的亮度，至少不是对每个像素等同地分析。人通过在脑子里分析这些特征并与先验知识

结合以完成对图像的解释过程。由于每个人所具有的先验知识不同，对同一幅图像的心理视觉冗余也就因人而异。心理视觉冗余从本质上说与前面几种冗余不同，它是与实在的视觉信息联系着的。只有在这些信息对正常的视觉过程来说并不是必不可少时才会被去除。去除心理视觉冗余数据能导致定量信息的损失，所以这个过程也常称为量化（也是一种由多到少的映射）。考虑到这里视觉信息有损失，所以量化是不可逆转操作，它被用于数据的有损压缩。根据心理视觉冗余的特点，可以采取一些有效的措施来压缩数据量。

尽管数据压缩的实践很早以来就广泛存在于人们生活当中，科学上对数据压缩的研究却始于 1948 年。针对通信的信源编码问题，Shannon 发表了“通信的数学理论”。在这篇文章里，作者从概率统计的角度详细讨论了输出信元为 N ($N=0, 1, 2, 3$) 阶平稳有限马尔科夫链时的无损压缩问题，并从理论上证明，信号的无损压缩倍率不可能高于信号的熵。在此基础上，1952 年 Huffman 提出了以他的名字命名的编码算法，解决了一阶马尔科夫序列的无损压缩问题。1963 年 Peter Elias 提出了比哈夫曼码更优的算术编码，但是第一个实际可用的算法直到 1976 年才出 Rissanen 和 Pasco 给出。1977 年 Ziv 等人基于字串复杂度分析，设计通用信源的 Lempel - Ziv 编码得到了和 Shannon 的熵编码相同的结果。此后，熵编码成为各种数据压缩技术的基础。从信息论的角度看来，数据压缩的问题转化为如何将高阶马尔科夫序列转化为一阶序列的问题。

设信号为 $F(x, y)$ ，则

$$F(x, y) = F_d(x, y) + F_s(x, y) + N(\sigma, \mu) \quad (1.3.1)$$

式中， $F_d(x, y)$ 为确定性信号； $F_s(x, y)$ 为马尔科夫随机信号； $N(\sigma, \mu)$ 为高斯噪声。

针对上述信号的混合模型，信号压缩的算法应该包含以下基本步骤：（1）信号逼近；（2）信号去噪；（3）熵编码。这种情况下，去相关的本质是二维曲面的分解与逼近，常用的方法有傅里叶分析或小波分析。通过某种正交变换可以极大地去除信号的相关；信号去噪是为了降低数据的熵值。噪声信号一般具有很高的熵值，混合在有效信号中对压缩编码的整体效率有很大影响，必须在熵编码之前剔除出去；经过上述处理以后，信号转化为一组随机信号，可进一步采用熵编码进行压缩。

综上所述，信号压缩的基本原理如图 1.3.1 所示，包含以下几个部分：去相关、量化（无损压缩没有量化这步骤）和熵编码。首先将信号当作确定性信号处理，采用 DFT、DCT（其实质仍为傅里叶分析）、DWT 变换等函数逼近的方法降低信号的相关性；然后对变换的系数进行量化处理，以去除信号的噪声，降低系数熵值；最后，采用信息论提供的方法进一步去除数据的冗余。

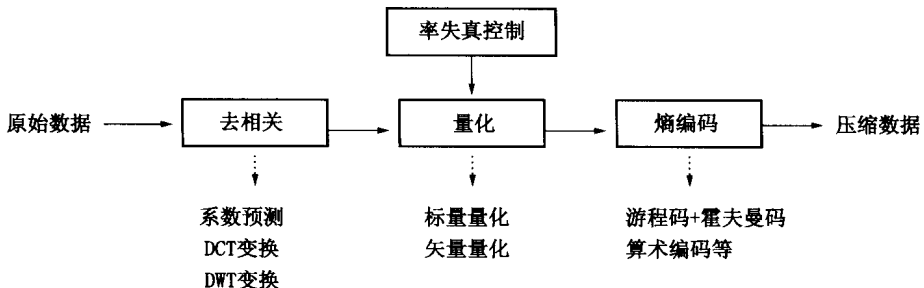


图 1.3.1 数据压缩原理框图

2 地震信号压缩中的变换方法

变换这个概念对数学家来说很熟悉，它在许多领域都是解决问题的一种强大的数学武器。变换的核心思想是将一个数学量从一种形式变成另一种形式。此时，该数学量可能看起来比较陌生，但却显现出了一些有用的特征。很多不能在变换前解决的问题在变换后就可以轻松地解决了，然后再将结果反变换为原来的形式。在压缩领域，变换也是一种不可缺少的重要工具。压缩领域中变换的主要目的是降低信号空间域或时间域表示中存在的相关性，使得变换后的数据变成低相关性的数据。理想的变换不仅能去相关，还可以将信号的主要能量集中于很少的几个系数上，这实际是信号的一种更紧凑的表示，从而有利于压缩的实现。本章简要分析地震信号的特点，重点探讨压缩中常用的两类变换方法——余弦变换和小波变换，并结合地震信号的特性，给出适用于地震信号压缩的表示。

2.1 概述

在广袤的平原、一望无际的戈壁滩和辽阔的海面上，有时我们会看到一座座井架和一口口油井。现代工业的血液就是从那里流出的。自然地，人们就会想，为什么这个地方有油，而其他地方没有呢？油井的位置是如何确定的？要回答这些问题，得从油气勘探说起。

要寻找深埋在地下几千米的油气资源，是一项非常具有挑战性的工作。人们经过不断的试验和总结，吸取和引用了许多其他学科的新技术、新理论，建立了一整套油气勘探的方法和技术体系。对于一些简单的构造地区，通过观测、研究裸露在地面的地层、岩石，对地质资料进行分析综合，了解一个地区有无生成油气和储存油气的条件，就可以给出对该地区的含油气评价。在有些情况下，由于地表结构的复杂，比如地表为松散沉积或沙漠覆盖的地区，或者海水覆盖的海洋地区，地面和海面上看不到岩石，地质法就难以取得好的效果，此时就需要应用物探方法，根据地质学和物理学的原理，利用电子学和信息论等领域的新技术，通过各种物理仪器，观测地壳上的各种物理现象，从而推断、了解地下的地质构造特点，寻找可能的储油构造的方法。

在勘探油气的各种物探方法中，地震勘探已成为一种最有效的方法。所谓的地震勘探就是利用人工方法引起地壳振动，如利用炸药爆炸产生人工地震，再用精密仪器记录下爆炸后地面上各点的震动情况。利用记录下来的地震反射波信号，推断地下地质构造的特点（陆基孟，1993）。那么人工地震为什么能查明地下地质构造呢？我们知道，当投一块石头到平静的水池里时，平静的水面就会出现一圈圈的波纹，向四面八方传播，形成“水波”。“水波”传到水池边或遇到障碍物时还会返回来，发生所谓的“波的反射”。地震勘探的原理与此十分类似，在地面上某点打井放炮后，爆炸产生的地震波向下传播。地震波遇到地层（速度与密度的乘积有差异）的分界面时，通常会发生反射和透射；同时另一部分地震波还会继续向下传播，碰到相似的地层界面后还会产生反射和透射，即一部分地震波的能量反射回地面，另一部分继续向下传播。与此同时，地面上精密的仪器把来自各个地层分界面的反射波引起地面振动的情况记录下来。然后根据地震波从地面开始向下传播的时刻和地层分界面反射波

到达地面的时刻，得出地震波从地面向下传播到达地层分界面，又反射回地面的总时间，再用别的方法测定出地震波在岩层中传播的速度，最后就可得到地层分界面的埋藏深度。如图 2.1.1 所示为地震勘探原理图。

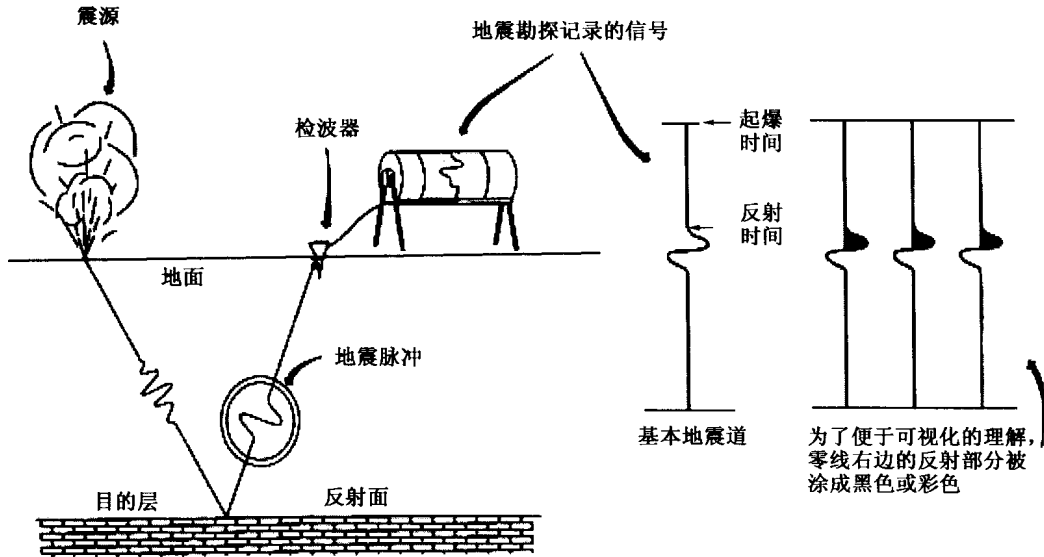


图 2.1.1 地震勘探原理示意图

沿着地面上的一条测线，一段一段地进行观测，对观测结果进行处理后，就可得到形象地反映地下岩层分界面埋藏深度起伏变化的资料——地震剖面图。在一个可能有油气的地区，布置多条测线，形成测线网，并在多条测线上进行这种观测之后，可得到地下地层起伏的完整概念，再综合其他物探方法和地质、钻井等各方面的资料，进行去伪存真，去粗取精，由此及彼，由表及里的分析、研究，就能查明可能储存油气的地质构造。概括地说，地震勘探，就是通过人工方法激发地震波，研究地震波在地层中的传播情况，查明地下地质构造，寻找油气田或其他矿物目标的一种物探方法。

2.2 地震信号分析

2.2.1 地震信号的信息量

信息论中信源模型可用随机过程描写。经过取样、量化后的地震数字信号信源是离散信源。离散信源支配着一个由 K 个符号 $\{a_1, a_2, \dots, a_K\}$ 组成的符号集，各符号出现的概率为 $p(a_1), p(a_2), \dots, p(a_K)$ ，且 $\sum_{k=1}^K p(a_k) = 1$ 。地震信源发出一个消息的过程可以理解为信源按时间进程顺序的从符号集选出符号形成一个符号序列 $s_1, s_2, \dots, s_n, \dots, s_N$ 的过程。在这个序列中， s_n 可以是符号集 $\{a_1, a_2, \dots, a_K\}$ 中的任何一个。

根据信息论，信源发出某一符号的信息量可以用这一符号出现的不确定性来描述。如果某个符号的出现可以比较准确地预测，那么这个符号就只含有少量的信息量；如果一个符号的出现难以预测，则这个符号就具有较大的信息量。换言之，一个符号出现的不确定性越

大，则它的信息量越大。因此，一个符号的信息量可以用其出现的概率进行度量，即

$$I(a_k) \propto \frac{1}{p(a_k)} \quad (2.2.1)$$

如果概率 $p(a_k)$ 用以 2 为底的对数表示，则信息量 $I(a_k)$ 的单位为比特，用公式表示为

$$I(a_k) = \log_2 \frac{1}{p(a_k)} = -\log_2 p(a_k) \quad (2.2.2)$$

如果信源相继发出的两个符号 a_i 和 a_j 是相互独立的，则从概率论可知，两个符号相继发出的概率为 $p(a_i)$ 和 $p(a_j)$ 。于是由公式 (2.2.2)，信源相继发出符号 a_i 和 a_j 的信息量为

$$I(a_i, a_j) = -\log_2 [p(a_i) \cdot p(a_j)] = I(a_i) + I(a_j) \quad (2.2.3)$$

上式表明，两个相继发出的、相互独立的符号 a_i 和 a_j 的信息量等于两个符号 a_i 和 a_j 各自信息量的和。

求一个离散、无记忆信源所发出的符号集中各符号的自信息量的统计平均值，便得到信源每个符号所传送的平均信息量为

$$H(A) = \sum_{k=1}^K p(a_k) \log_2 [1/p(a_k)] \quad (2.2.4)$$

$H(A)$ 被成为信源的熵。所以熵可被认为是系统不确定程度的度量。当所有事件具有等概率时，事件将在某一时间出现的不确定性为最大，这时候 $H(A)$ 亦最大，而对确定事件而言， $H(A)$ 为零。

2.2.2 地震信号的相关性分析

地震勘探过程中采集的地震信号，反映了地下地质信息。在垂直（时间）方向上，由于岩层沉积的渐变性，相邻采样信号具有比较强的相关性；在水平方向上，相邻道之间由于空间上比较近，所以道与道之间也存在相关性。

地震信号的接收过程可看作一个随机过程。一维离散随机过程 $f(x_m)$ 的自相关函数定义为

$$R(\Delta x) = \frac{1}{M} \sum_{m=1}^M f(x_m) f(x_m + \Delta x) \quad (2.2.5)$$

式中， Δx 为 x 方向的间隔距离。

经常使用归一化自相关函数来度量信号间的相关性。一维归一化自相关函数定义为

$$r(\Delta x) = \frac{\sum_{m=1}^M [f(x_m) - \bar{f}] \times [f(x_m + \Delta x) - \bar{f}]}{\sum_{m=1}^M [f(x_m) - \bar{f}]^2}, \quad \text{其中 } \bar{f} = \frac{1}{M} \sum_{m=1}^M f(x_m).$$

图 2.2.1 为原始单炮地震数据的波形剖面图；图 2.2.2 为其水平和垂直方向的相关性分析。从图中可以看出，对原始单炮数据而言，在时间（垂直）方向上相关性比较强，而道与道之间由于初至波的时间差，水平相关性较弱。图 2.2.3 为叠加处理成果地震数据的波形剖面图；图 2.2.4 为其水平和垂直方向的相关性分析。从图中可以看出，对成果数据而言，由

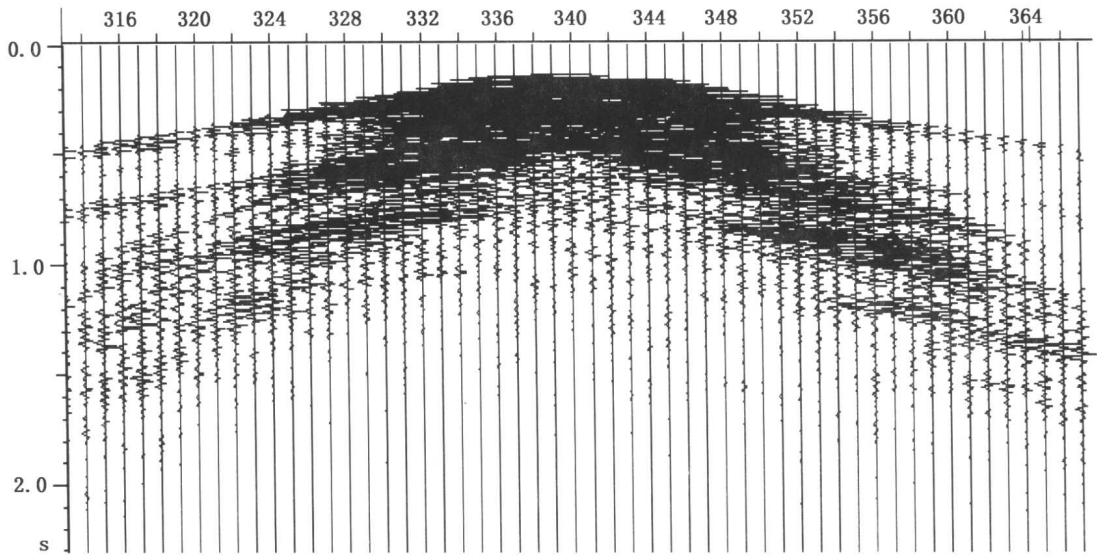


图 2.2.1 原始单炮地震数据

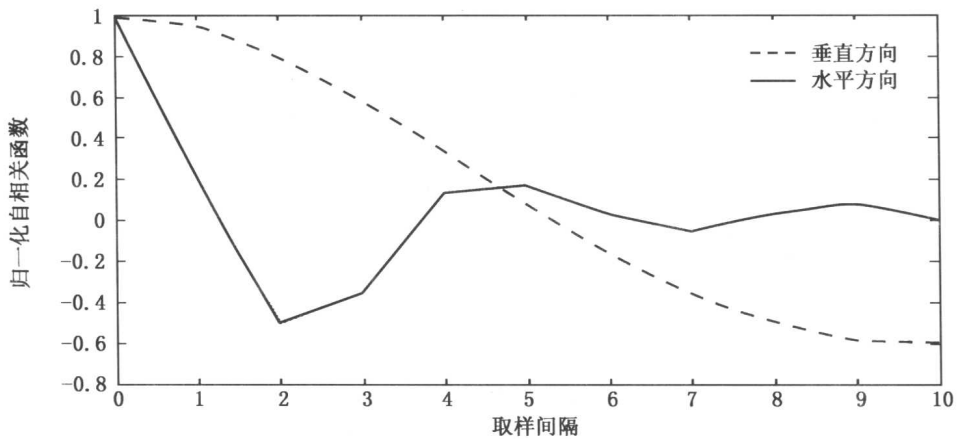


图 2.2.2 原始单炮地震信号的相关性分析

于进行了叠加处理，消除了初至波的影响，水平和垂直方向相关性都较强。

2.2.3 地震信号的功率谱分析

地震信号的功率谱可以经傅里叶变换在频率域测量，也可以通过在空间域测量的自相关函数间接计算得到。信号分析理论中的维纳—辛钦定理证明，信号的自相关函数与它的功率谱互为傅里叶变换对，因此，功率谱密度 $S(f)$ 可通过下式由其自相关函数 $R(\tau)$ 的傅里叶变换得到，即

$$S(f) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} R(\tau) e^{-j2\pi f\tau} d\tau$$

图 2.2.5 为某道地震信号的时域表示，图 2.2.6 为其傅里叶分析图。上述实验结果表明，地震信号的频谱在 $0 \sim 100\text{Hz}$ 范围内会有一个峰值，并且在这个范围内功率谱数值较

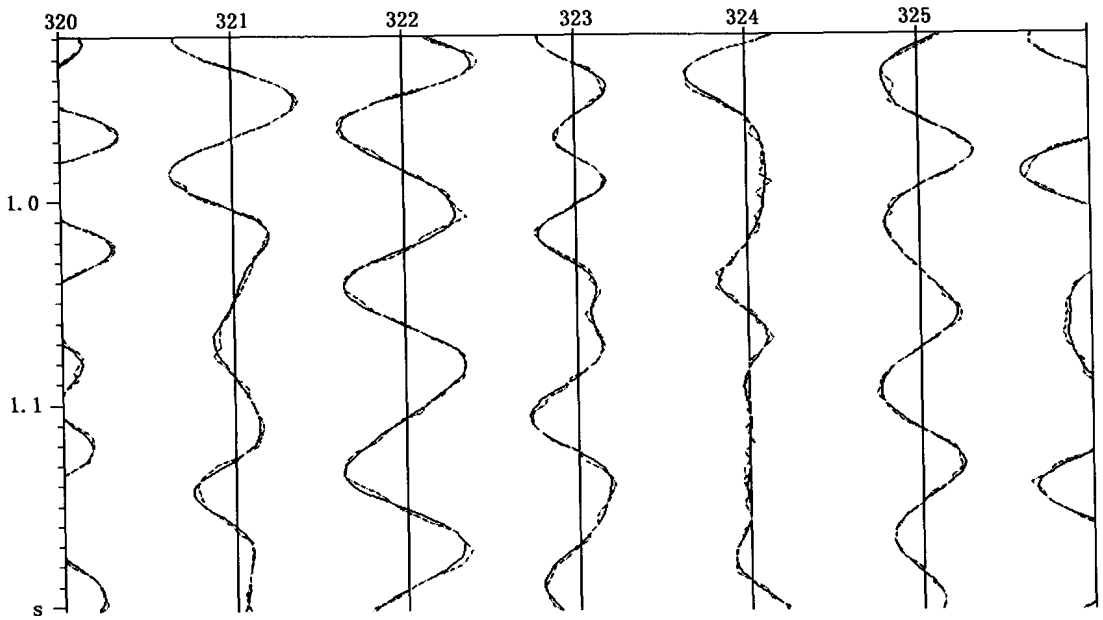


图 2.2.3 叠加处理成果地震数据

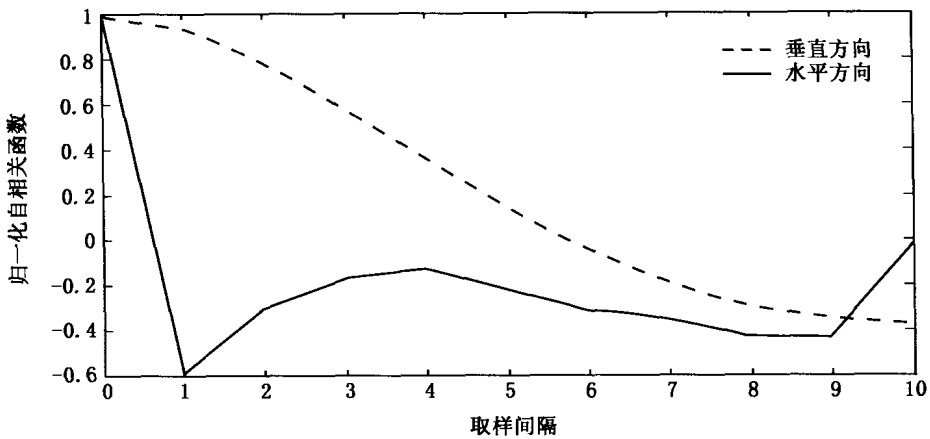


图 2.2.4 叠加处理成果地震数据相关性分析

大；在频率超过 100Hz 时，功率谱的值已低于 -50dB (图 2.2.6)。这些特性表明，地震信号的大部分能量集中在低频部分，高频成分较少。

2.2.4 地震信号的模型分析

如果能建立一个简单的、精确的地震信号统计模型，那么这会对地震信号处理带来很多的好处。对压缩来说，利用可靠的统计模型，再加上可靠的概率分布测量，利用信息论中的率失真理论，就可能达到压缩编码码率极限 (张春田, 2006)。对于预测编码，利用统计模型可以优化预测器和量化器的设计，并且可以实现与统计模型完全相匹配的熵编码，如霍夫曼编码或算术编码；利用统计模型可以实现最优变换编码，在不需要实时测量信号自相关函数的情况下，就可以推导变换的核函数等。