1
Introdução

Os dados sísmicos guardam informações colhidas sobre o subsolo de uma região e são utilizados pelos geofísicos para analisar e estudar tal região de terreno. Os objetivos variam desde o estudo sobre ocorrências de atividades sísmicas como vulcões e terremotos até a identificação de possíveis poços de combustíveis fósseis.

Para obter estes dados, explosivos e sensores são espalhados sobre uma área. A detonação dos explosivos gera vibrações cujas ondas viajam em profundidade e são refletidas pelas camadas de terreno. Os sensores fazem leituras da intensidade das vibrações em intervalos regulares de tempo. As informações de cada sensor são depois cruzadas e processadas gerando, para cada ponto do plano, um vetor contendo a amplitude da vibração para uma seqüência de intervalos de tempo. Este vetor é chamado de traço sísmico.

Em média, o intervalo de tempo entre cada leitura é 4ms. Se uma amostragem durar de quatro a dez segundos temos, para cada traço, entre mil e dois mil e quinhentos valores. Para uma área coberta com 2.000x2.000 sensores temos 4.000.000 traços e um dado de aproximadamente 30Gb assumindo 4 bytes por amostra. A Tabela 1 mostra exemplos de dimensões de alguns dados.

Dado	R	C	F	P	B	I
Dimensões	126 x 70 x 100	126 x 161 x 184	255 x 256 x 256	401 x 451 x 2750	560 x 738 x 2250	1091 x 2801 x 750
Número de Amostras	882.000	3.732.624	16.711.680	497.340.250	929.880.000	2.291.918.250
Tamanho (bytes)	3.528.000	14.930.496	66.846.720	1.989.361.000	3.719.520.000	9.167.673.000

Tabela 1: Exemplo de dimensões e tamanho de arquivos sísmicos.

Devido ao tamanho que estes dados podem atingir sua manipulação deve ser bem planejada, caso contrário haverá um consumo proibitivo de recursos computacionais. Sua transferência será custosa em tempo e seu

armazenamento, muitas vezes, inviável para uma máquina pessoal típica da atualidade.

O objetivo deste trabalho consiste em apresentar propostas para manipulação dos arquivos de dados sísmicos de forma a minimizar os problemas conseqüentes de suas dimensões. Nosso foco é diminuir o tempo de transferência dos dados de disco para a memória durante uma operação típica de aplicativos de geofísica; e diminuir o custo com transferência e armazenamento destes arquivos.

1.1. Um novo formato para armazenar dados sísmicos

Muitos aplicativos que possibilitam a manipulação de dados sísmicos provêem maneiras de selecionar regiões para visualização. Estas regiões variam de fatias paralelas às faces do volume, fatias diagonais, sub-volumes ou amostras cujos valores satisfazem algum critério conforme é exemplificado nas Figura 1 e Figura 2 a seguir. Normalmente, nestes mesmos aplicativos, os usuários podem trabalhar com mais de uma seleção simultaneamente, bem como trocá-las à vontade. Devido a esta liberdade, dizemos que as amostras são selecionadas para uso de forma aleatória.

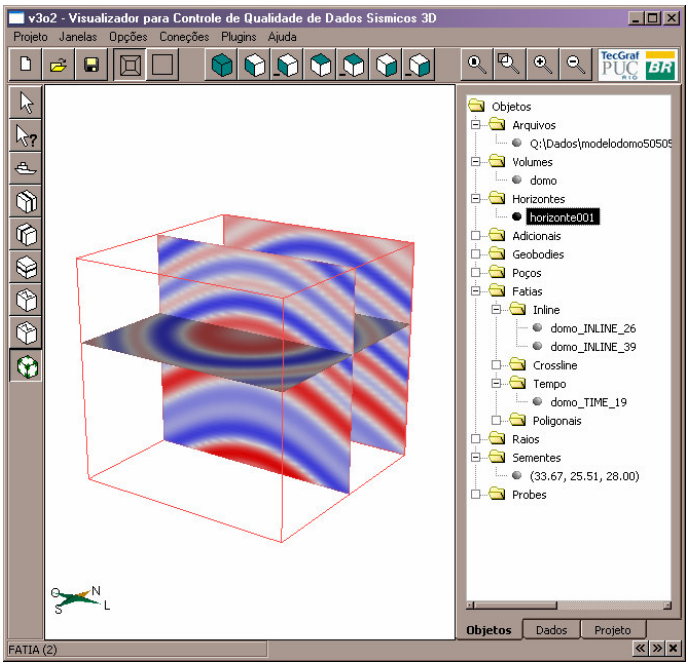

Figura 1: Aplicativo sendo utilizado para ver fatias de um dado.

Computadores modernos trabalham com um modelo de memória em camadas partindo desde a *cache* do processador até disco rígido. Quanto mais próxima do processador na hierarquia estiver a memória, mais rápido é o seu acesso, porém a capacidade de armazenamento decresce no mesmo sentido. Além disso, as arquiteturas atuais mantêm o controle do conteúdo das *cache*s para o processador em tempo de execução, deixando a programação com acesso apenas a *RAM* e ao disco. A *RAM,* cujo nome significa *random access memory,* é ideal para acessos aleatórios. Já o disco funciona melhor para acessos em blocos contínuos. Neste momento fica clara a dificuldade apresentada pela dimensão dos dados sísmicos – as máquinas atuais de 32bits não podem possuir mais de 4Gb de memória e as máquinas de 64bits ainda estão começando a ganhar seu espaço no mercado. Mesmo onde já se trabalha com máquinas de 64bits o custo financeiro da memória é alto e não é comum encontrar máquinas cuja disponibilidade de *RAM* seja compatível com a dimensão dos dados que, comumente, apresentam muitas dezenas de gigabytes, não sendo tão raro apresentarem poucas centenas. Esta impossibilidade de armazenar o dado por completo em memória é um problema para o trabalho com os mesmos.

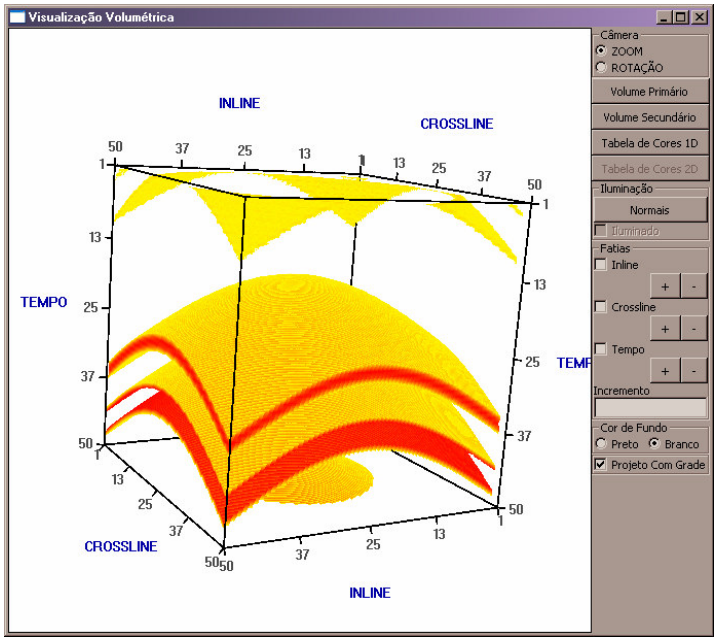

Figura 2: Aplicativo de visualização volumétrica sendo utilizado para ver amostras cujos valores encontram-se entre um intervalo escolhido.

Dentre algumas possibilidades para contornar o problema consideremos a opção de sub-amostragem regular. Em vez de manter todo o volume em memória carrega-se apenas uma a cada *n* amostras em cada direção. Os valores e imagens dos pontos não carregados são gerados por interpolação. Nesta técnica, dependendo do tamanho do dado e da memória disponível, o erro associado pode ser muito grande e detalhes importantes podem ser ocultados.

Outra solução é carregar um corte volumétrico do dado de cada vez e fazer uma série de trabalhos independentes. Neste caso, um sub-volume é carregado para a memória e o trabalho fica limitado àquela região. Imaginemos a situação em que um geofísico precisa analisar fatias paralelas às faces do volume. Ao adotar esta solução de trabalhos independentes ele terá que analisar partes da fatia por vez, tantas vezes quantas forem o número de divisões. Esta análise em etapas pode, por exemplo, comprometer a percepção de eventos que ocorrem nas bordas das sub-fatias, como também apresentar outros problemas de continuidade. Este problema não ocorre na análise da fatia como um todo.

O problema apresentado nesta última solução é conseqüência do fato que a parte carregada para a memória não equivale à parte com a qual se deseja trabalhar, no caso, a fatia, que por sua vez poderia ser pequena o suficiente para caber na memória. Uma variação mais inteligente desta solução é carregar uma sub-seleção que inclua totalmente a região de trabalho. O problema agora fica apenas na questão do tempo. Com o dado todo armazenado em disco, a troca de seleção pode ser muito custosa. Uma de nossas propostas, então, visa permitir que este trabalho seja feito minimizando o custo relativo à transferência de disco para memória. Em nossos experimentos identificamos como principal razão para tal custo a maneira como os dados são organizados. Aqui mostramos como diferentes organizações e estratégias de leitura afetam este tempo e apresentamos uma organização direcionada para o tipo comum de uso de dados sísmicos. Com o esquema proposto conseguimos diminuir consideravelmente o tempo das leituras mais prejudicadas no esquema atual sem apresentar perda significativa na eficiência de outras leituras igualmente importantes.

1.2. Compressão de dados sísmicos

A dimensão dos dados também dificulta a transferência e o armazenamento. Este problema pode ser abordado com técnicas de compressão. O estudo sobre compressão de dados sísmicos não é novidade na

literatura. Por se tratar de dados provenientes da leitura de reflexões de vibrações, a grande parte dos trabalhos de compressão se baseia em teorias de ondas. As técnicas mais comuns são decomposições e transformadas *wavelets* [1,2,7], e de co-seno [9,10]. Em [10] são apresentadas transformadas de seno e/ou co-seno semi-adaptativa para substituir as transformadas uniformes. No trabalho de Averbuch, A. et Al. [1] são comparados estes métodos com foco na relação entre as taxas de compressão com as velocidades dos procedimentos. Outras muitas versões e adaptações destas transformadas são constantemente estudadas e publicadas no universo de geofísica e processamento de sinais, incluindo imagens.

No aplicativo v3o2 uma compressão muito simples é utilizada para parte dos procedimentos. O método em questão divide o espaço de valores em 256 intervalos uniformes montando uma tabela de quantização. O valor de cada amostra é, então, substituído pelo índice do valor da tabela que mais se aproxima de seu valor original. A escolha de 256 intervalos permite que esta indexação seja representada com um byte. O arquivo de índices é, então, quatro vezes menor que o original uma vez que este utiliza ponto flutuante de 32 bits para cada valor. Este aplicativo faz uso desta compressão na etapa de visualização onde cada índice é associado a uma cor. Assim como o v3o2 muitos outros aplicativos, por exemplo gOcad, Geoprobe®, Promax® e Petrel®, fazem associação das cores seguindo o mesmo princípio. Inclusive o mesmo número, 256, de índices é utilizado pois cada um deles representa uma tonalidade de cinza.

Embora esta quantização seja largamente aceita nas aplicações de visualização é muito comum que estes mesmos aplicativos mantenham os originais do dado para outras atividades como, por exemplo, procedimentos matemáticos delicados. Deste contexto surgiu a inspiração para este trabalho onde resolvemos nos distanciar da compressão sísmica tradicional por ondas e abordar o problema com técnicas de agrupamento. Aqui propomos uma nova técnica de escolha dos grupos baseada em minimização do erro médio e comparamos nossa proposta com a quantização por intervalos uniformes usando como medida o erro e a capacidade de compressão. Também observamos o comportamento destas medidas em função do número de grupos. Embora ambas quantizações apresentem curvas da relação erro por capacidade de compressão muito semelhantes, a estratégia aqui apresentada consegue seus resultados utilizando sempre um número significativamente menor de grupos.

Para concluir esta etapa do estudo analisamos maneiras diferentes de codificar o dado após o agrupamento tentando minimizar seu tamanho. A coerência espacial, característica de dados sísmicos, foi a base das técnicas apresenta a codificação. Em nosso experimentos, a solução final montada juntando agrupamento com aproveitamento da coerência lateral conseguiu compressões variando de 7% a 25% dependendo do erro associado.

1.3. Organização da dissertação

A dimensão dos dados, conforme apresentado, é um problema que afetas várias etapas de sua manipulação. Com base nestas etapas dividimos o trabalho em duas partes. No capítulo 2 apresentamos a reorganização dos dados onde analisamos como seu armazenamento influi no tempo de transferência para a memória de diferentes sub-seleções do mesmo. Apresentamos a condição atual, a motivação para a proposta, nossos resultados e finalmente possíveis otimizações e variações das mesmas.

Em seguida, no capítulo 3 temos uma proposta para compressão dos dados. Ao longo das seções deste capítulo acompanhamos as várias etapas da compressão e estudamos seus problemas e diferentes maneiras de abordá-las.

O capítulo 4 é destinado à conclusão onde analisamos o trabalho desenvolvido, apresentamos um levantamento de como combinar as técnicas de compressão estudadas com a nossa proposta de reorganização dos dados e descrevemos alguns trabalhos futuros.