

## 5 Resultados

Os resultados da aplicação do método proposto estão estruturados de forma a evidenciar:

1. A convergência numérica do algoritmo proposto nesta tese para a mediana e para a média de grupos não simétricos;
2. O desempenho na classificação de grupos do algoritmo proposto nesta tese na comparação com outros algoritmos, em particular os citados na literatura;
3. A aplicação em dados reais.

Os resultados expostos nesta tese foram obtidos com dois tipos de dados: dados artificiais e dados reais oriundos da Bacia de Campos no Brasil fornecidos pela Petrobras.

### 5.1. Dados artificiais

#### 5.1.1. Convergência à mediana e à média com dados artificiais

O algoritmo proposto foi aplicado em quatro diferentes tipos de distribuições com o objetivo de avaliar a convergência à mediana e à média. As distribuições foram: misturas de distribuições normais, distribuição exponencial, distribuição aleatória e distribuição aleatória enxertada com “*outliers*”. Para cada grupo de dados, gerado artificialmente, dois testes foram realizados. No primeiro teste, com  $h$  constante, foi avaliada a convergência à mediana. No segundo teste com  $h$  variável, foi avaliada a convergência à média. O valor de  $h$  variável foi calculado por  $h = e^{\exp(-\frac{p}{P})}$  onde  $p$  indica a variação da época no processo e  $P$  a quantidade total de épocas determinada pelo usuário.

Todos os testes com dados artificiais foram realizados várias vezes (mínimo de 5). Cada ciclo englobou a geração de números com os mesmos parâmetros e o processamento do algoritmo após. Os resultados listados abaixo expressam um dos testes. Não foi registrado caso de exceção nos resultados esperados.

### Misturas de distribuições normais

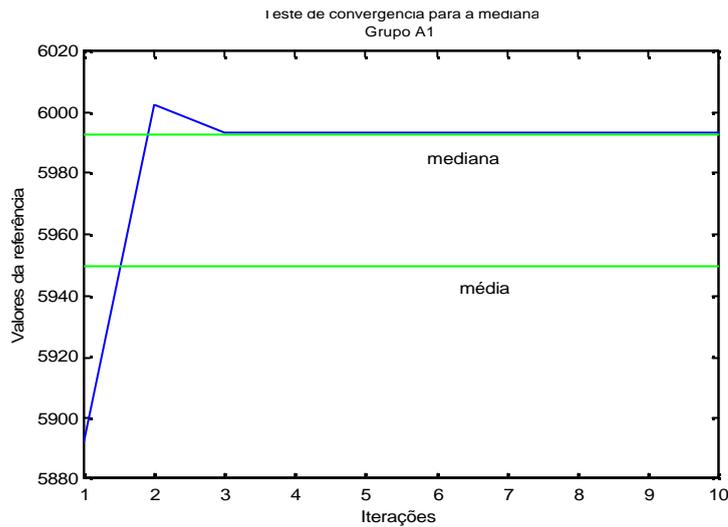
O conjunto A1 foi gerado com a mistura de quatro distribuições normais, cada qual com o mesmo número de elementos. As tabelas abaixo mostram a composição das misturas e os parâmetros do grupo A1 formado.

Misturas normais	Quantidade de elementos	Mistura
A1	10000	$0.25N(5900,100)+0.25N(5900,90)+0.25N(6000,10)+0.25N(6000,5)$

	Média	Mediana	Desvio Padrão
A1	5949	5993	84.11

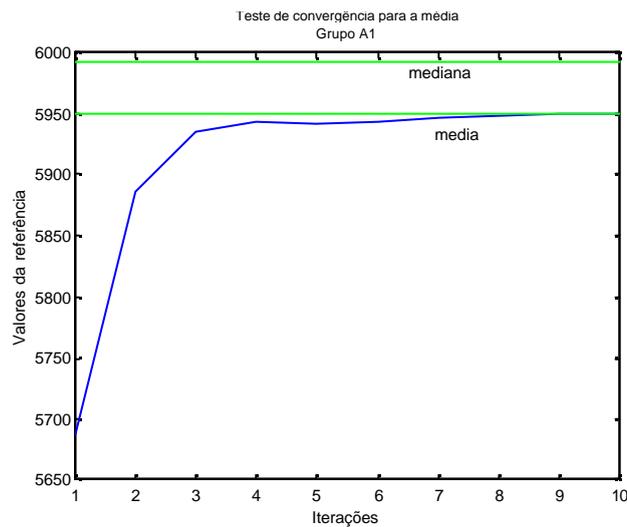
### Mediana

O processo com A1 utilizou  $c = 1300$  e  $h = 1$ . O valor da referência de 5993 foi atingido antes de 10 iterações. O gráfico abaixo mostra a convergência da referência à mediana.



### Média

O processo com A1 utilizou  $c = 0$  e  $h =$  variável. O valor da referência de 5993 foi atingido antes de 10 iterações.



### Distribuição exponencial

O grupo A2 foi gerado com 10000 elementos com a seguinte lei de formação:

1. Foram gerados 10000 números aleatórios entre 0 e 1;
2. A expressão para o conjunto final foi:  $y_i = -\log(x_i) * 10$  onde  $i = \{1, \dots, 10000\}$  e  $x_i$  correspondem aos números aleatórios gerados no item 1. No cálculo de  $y$ , o multiplicador 10 provoca  $s > 1$ .

As características dos dados gerados são as seguintes:

Dados exponenciais	Média = 9.88	Mediana = 6.85	Desvio = 9.90
-----------------------	--------------	----------------	---------------

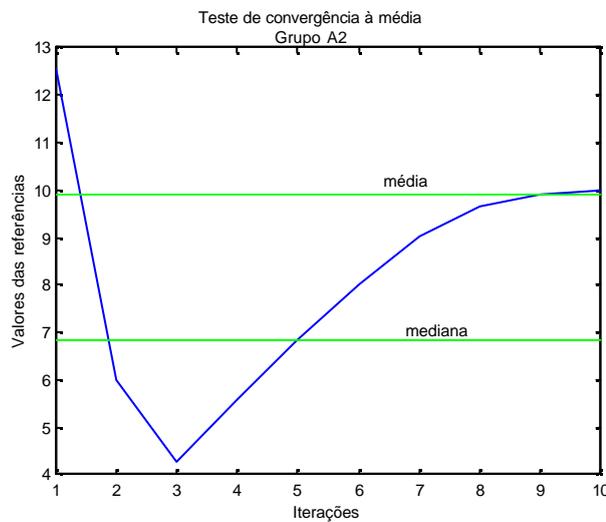
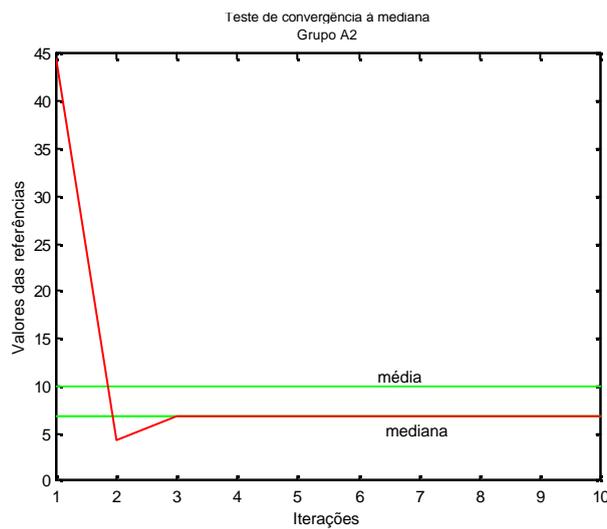
**Mediana**

O processo do algoritmo proposto após 10 épocas mostrou um valor de referência para o grupo de 6.89 com  $c = 48$  e  $h = 0.2$ . Neste teste a constante de não linearidade foi utilizada no ponto máximo.

**Média**

No que concerne o processo que procurou induzir a referência à média, foi encontrado um valor de 9.93 antes de 10 épocas.

Os gráficos abaixo mostram o desempenho do processo.



### Distribuição aleatória

O grupo A3 foi gerado com 10 elementos aleatórios segundo a fórmula:  $y_i = rand(m) * 100$  onde  $m = 10$ . As características dos dados gerados são as seguintes:

Dados aleatórios	Média = 27.09	Mediana = 18.80	Desvio = 24.96
------------------	---------------	-----------------	----------------

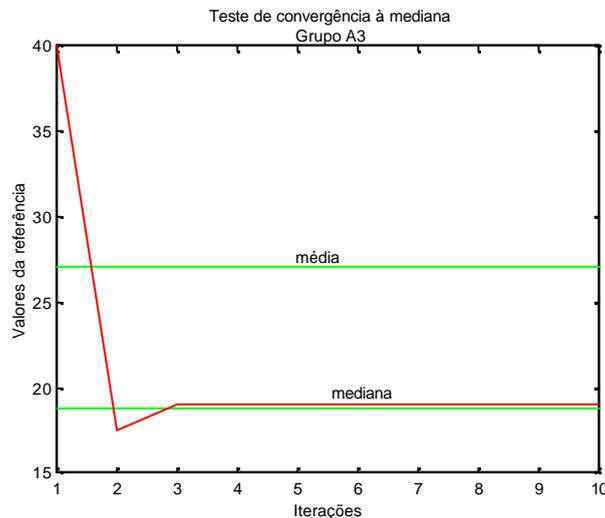
Duas são as razões para a pequena quantidade deste grupo. A primeira diz respeito à necessidade de testar o algoritmo com poucos elementos. A segunda está associada ao fato da facilidade de conseguir grupos com defasagem mais significativa entre média e mediana. Com efeito, para grande conjuntos de elementos a tendência da distribuição aleatória é a de aproximação entre média e mediana.

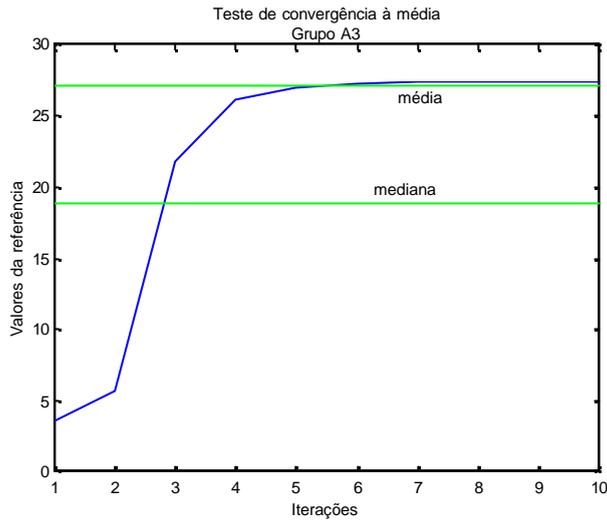
### Mediana

O processo do algoritmo proposto após 10 épocas mostrou um valor final para a referência do grupo igual a 19.03 com  $c = 155$  e  $h = 0.61$ .

### Média

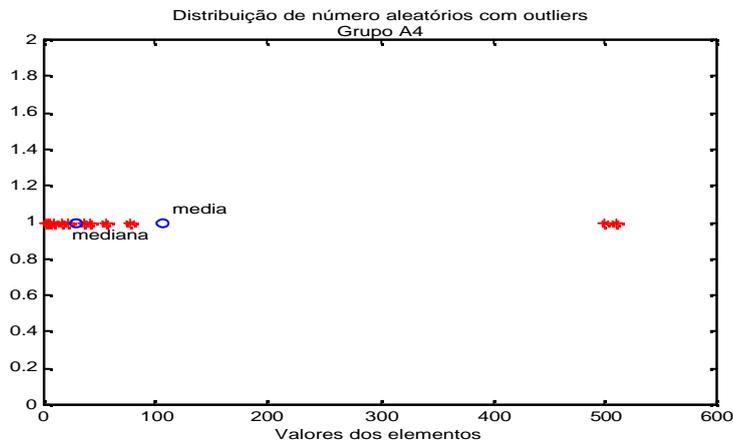
Após dez épocas, o valor final da referência foi de 27.37. Os gráficos abaixo mostram o desempenho do processo.





**Distribuição aleatória com “outliers”**

A partir do mesmo grupo de dados aleatórios do teste anterior, dois outros elementos foram incluídos no conjunto de testes: 500 e 505. O novo conjunto foi chamado de A4.



As características dos dados gerados são as seguintes:

Dados aleatórios com “outliers”	Média = 106.32	Mediana = 29.11	Desvio = 186
---------------------------------	----------------	-----------------	--------------

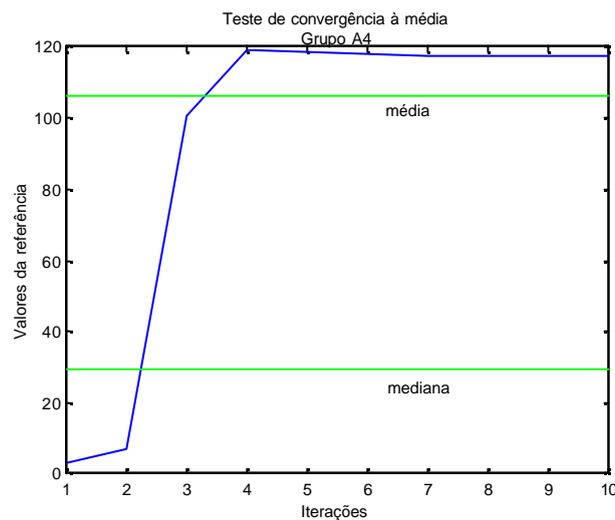
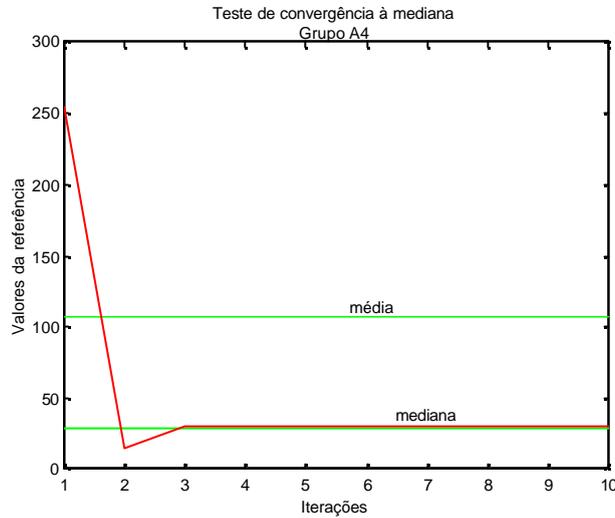
**Mediana**

Após 10 épocas, o valor final da referência do grupo foi de 30.20 com  $c = 6000$  e  $h = 0.9$ . Observa-se que o processo convergiu para a mediana

utilizando uma constante de não linearidade de valor relativamente baixo com um atenuador praticamente máximo.

### Média

O valor final da referência do grupo orientado à média foi de 117.



## 5.1.2. Análise de desempenho dos algoritmos

### 5.1.2.1. Geração dos dados para os experimentos controlados

O conjunto de dados artificial B1 foi gerado para a avaliação de desempenho. O grupo B1 é formado de misturas de distribuições normais sintetizando três camadas de dados sísmicos.

A formação das misturas que geraram B1 é descrita abaixo.

B1	Quantidade de elementos	Mistura
Camada 1	4800	$0.25N(5900,100)+0.25N(5900,90)+0.25N(6000,10)+0.25N(6000,5)$
Camada 2	4400	$0.25N(6100,100)+0.25N(6100,90)+0.25N(6200,10)+0.25N(6200,5)$
Camada 3	5200	$0.25N(6300,100)+0.25N(6300,90)+0.25N(6400,10)+0.25N(6400,5)$

Os paramentos de B1 são:

	Quantidade de elementos	Média	Mediana	Desvio padrão
B1	14400	6155	6190	186.95

Após a geração de B1 seus elementos foram aleatoriamente misturados.

Os testes foram avaliados através de um percentual de erro da classificação.

O percentual de erros  $e$  é calculado através da formula:

$$e = \frac{\text{quantidade classifica da errada}}{\text{quantidade total de elementos no grupo}} * 100$$

### 5.1.2.2.

#### Desempenho do algoritmo proposto

No que concerne o algoritmo proposto, a determinação das constantes  $c$  e  $h$  devem preceder os processos. Uma maneira prática de achar estas constantes é, a partir do desvio padrão de B1, determinar  $h$  para diversas constantes  $c$  encontrando qual é o processo que tem menor percentual de erro. O valor de  $c = \frac{s^2}{4}$  e  $h=05$  (valores centrais dos limites de cada constante) são indicados para o início da determinação do valor destas constantes. Para B1 a determinação de  $c$  e  $h$  é a seguinte:

1. O desvio padrão de B1 é  $s = 186.45$  o que implica nos limites para  $c$  iguais a  $0 \leq c \leq 17382$ .

2. A procura pelo melhor valor da constante  $h$  foi realizada a partir do

valor  $c = \frac{s^2}{4} = 8690$ . O percentual de erros para os processos com

os valores de  $c = 8000, 8690, 9000$  são mostrados na tabela abaixo.

	$h = 0.5$	$h = 0.6$	$h = 0.7$	$h = 0.4$	$h = 0.01$
$c = 8000$	10.13%	10.18%	11.56%	33.49%	41.47%
$c = 8690$	10.11%	10.19%	11.50%	10.18%	27.14%
$c = 9000$	10.18%	10.17%	11.56%	10.14%	27.26%

O melhor desempenho de classificação do algoritmo proposto para B1 é para  $c = 8690$  com  $h = 0.5$ . Observa-se também que na medida que  $h$  cresce ou diminui para o valor de  $c$  com menor percentual de erro encontrado (10.11%) os erros crescem.

A partir do valor de  $h = 0.5$  os limites de  $c$  foram testados. A tabela abaixo mostra diversos valores de  $c$  para a mesma constante  $h$ .

Valores de $c$	18000	17000	10000	9000	8690	8000	6000	0
% erro	90.04	44.90	10.18	10.18	10.11	10.13	10.60	15.93

Observa-se que para valores superiores ao limite máximo ( $c=17382$ ) o percentual de erro é de tal magnitude que o processo deve ser desconsiderado. A quantidade de épocas utilizada para obter os resultados foi de 200.

### 5.1.2.3. Resultados comparativos

Etapa seguinte é a de determinar o percentual de erro de outros algoritmos.

Os algoritmos testados foram:

- O algoritmo proposto com  $h$  variável e  $c = 0$  (indução à média);
- Som de Kohonen;
- O algoritmo de cálculo da mediana definido por Georgakis et al. (2004).

O objetivo deste tipo de teste é saber se o algoritmo proposto com indução à mediana classifica dentro dos padrões dos algoritmos existentes. Os testes foram realizados com 5000 épocas. A tabela abaixo mostra os resultados.

	Algoritmo Proposto $c=0$ $h$ variável	Kohonen	Georgakis et al. 2004	Algoritmo Proposto $c=8690$ $h=05$ 200 épocas
Percentual de erro	15.61%	13.43%	16.63%	10.11%

O algoritmo proposto classifica dentro dos padrões dos algoritmos existentes e, neste teste, obteve o menor percentual de classificação.

## 5.2.

### Dados sísmicos reais

O segundo tipo de dado é um conjunto real de sinais sísmicos (amplitude). Os dados sísmicos utilizados nesta tese foram fornecidos pela Petrobras. O conjunto engloba 772800 pontos de atributos amplitude sísmica distribuída em igual número (96600) dentre 8 camadas. Cada atributo é identificado pelas suas coordenadas de posição  $x$ ,  $y$ ,  $z$  e camada.

O conjunto de dados tem as seguintes características:

	Média	Mediana	Desvio Padrão
Dados reais	6446.7	6373	465.94

A finalidade do teste é a de avaliar, através de mapas, a diferença entre a classificação de um algoritmo usualmente utilizado para a classificação de dados sísmicos (SOM de Kohonen) e o algoritmo proposto no que concerne o atributo adquirido. Outro item a ser avaliado é como o algoritmo proposto permite o processo para a determinação de estruturas fantasmas.

### 5.2.1.

#### Erro dos processos

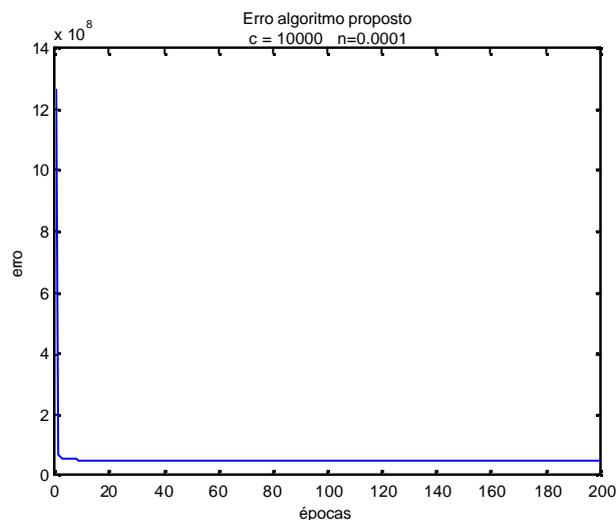
Os erros médios quadráticos dos processos do algoritmo proposto estão relacionados na tabela 3.

Valor de $c$	10000	10000	10000	10000	20000	20000
Valor de $h$	0.5	0.001	0.0001	0.00001	0.5	0.0001
Erro quadrático	97.23	62.30	62.06	64.13	561.52	69.73

Tabela 3 – Valores de erros de processos

O processo que obteve o menor erro para  $c$  igual ou superior a 10000 foi escolhido para a geração dos mapas. Há uma perda na qualidade da visualização do mapa para valores de  $c$  menores do que 10000. As constantes para o processo do algoritmo proposto foram  $c = 10000$  e  $h = 0.0001$ .

Os processos com os dados reais fornecidos convergiram rapidamente para resultados que permitiram mapas com melhor qualidade de visualização de estruturas. Não mais do que 300 épocas foram utilizadas nos processos com o algoritmo proposto. O gráfico abaixo mostra um exemplo da rápida convergência dos processos. Os processos com o algoritmo SOM de Kohonen utilizaram 5000 épocas.



### 5.2.2. Classificação do atributo adquirido

As classificações foram realizadas em nove grupos. Como descrito no capítulo 1, nove grupos é uma quantidade que permite a visualização de formas estruturais, estratigráficas e o isolamento de ruídos.

Para cada processo, junto com os mapas, são calculados os seguintes dados estatísticos dos grupos: valores mínimos e máximos, média, desvio padrão e quantidade de elementos classificada em cada grupo.

Os mapas são produzidos em 8 camadas. No entanto, para a comparação dos resultados entre os dois algoritmos somente uma camada é escolhida em cada mapa. Um exemplo de um mapa completo resultante do processo do algoritmo SOM de Kohonen é mostrado na figura 12, ao lado, o mapa do algoritmo proposto é mostrado na figura 13. As figuras 14 e 15 detalham a camada 3 de cada processo.

Cada classificação, como foi informado, gera uma tabela de dados que, para o SOM de Kohonen esta descrito na tabela 1 e para o algoritmo proposto na tabela 2.

Observa-se nos mapas que o algoritmo proposto, com as constantes mencionadas, determina um maior detalhamento de estruturas do que o algoritmo SOM de Kohonen.

É possível depreender da observação dos dados das tabelas:

1. Os desvios padrões do algoritmo Som de Kohonen dos grupos 2 ao 8 são baixos. Esta é uma característica deste algoritmo que busca propor grupos com os menores valores possíveis dos desvios padrões. Os grupos 1 e 9 aglutinam os elementos que não puderam ser classificados neste critério. A falta de detalhamento de estruturas geológicas nos resultados do algoritmo SOM de Kohonen é em função destes dois grupos;
2. Os desvios do algoritmo proposto mostram-se mais homogêneos, exceção feita ao grupo 1 que possui reduzida quantidade de elementos;

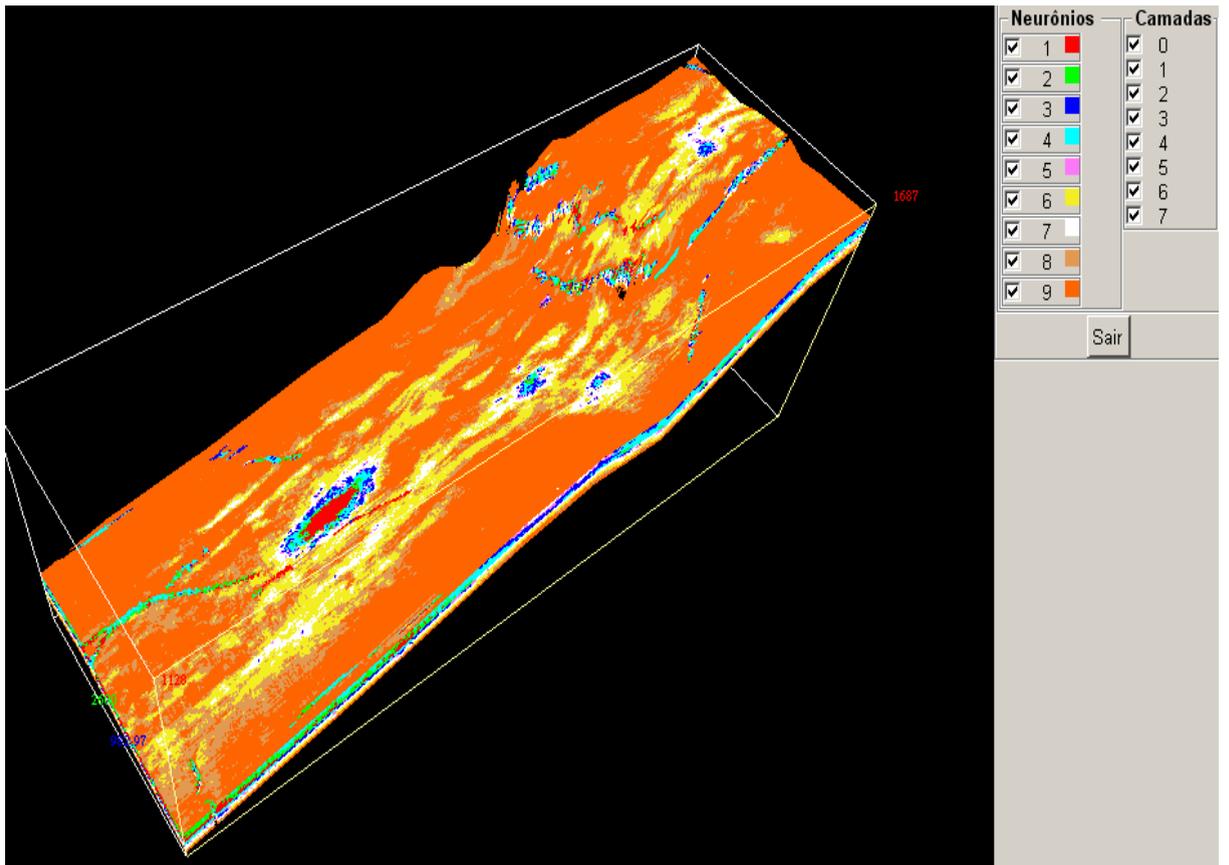


Figura 12 – Mapa sísmico – Kohonen

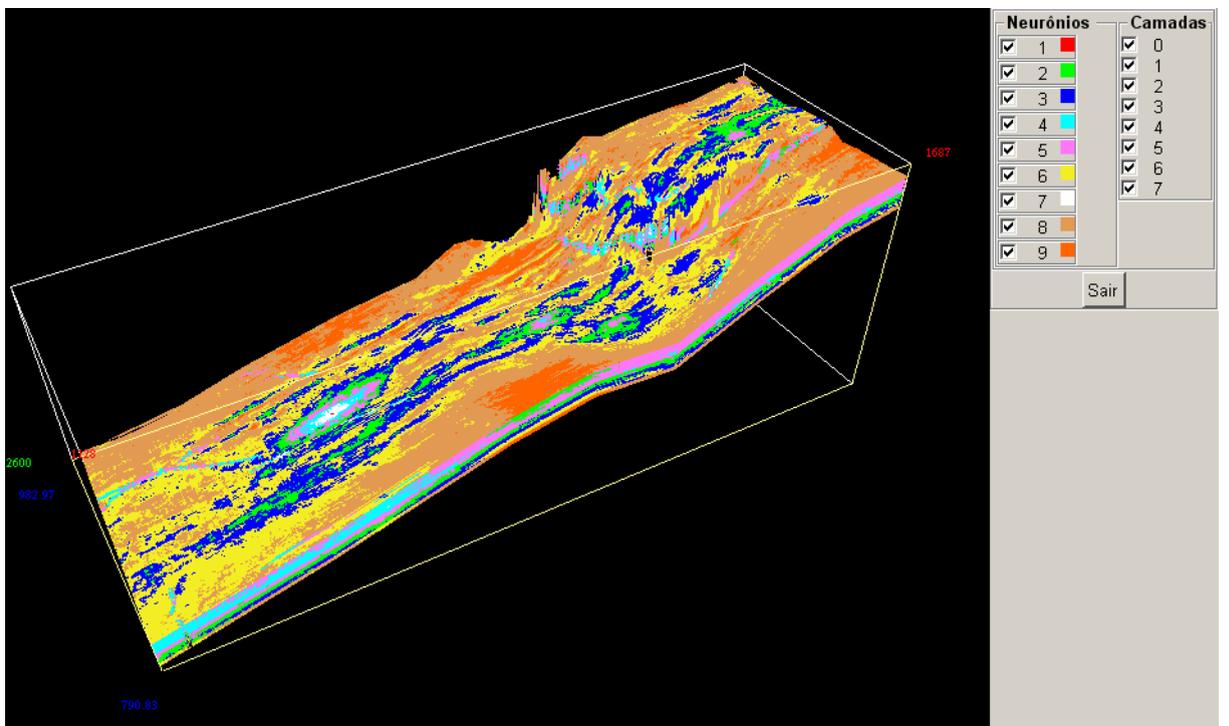


Figura 13 – Mapa sísmico - proposto

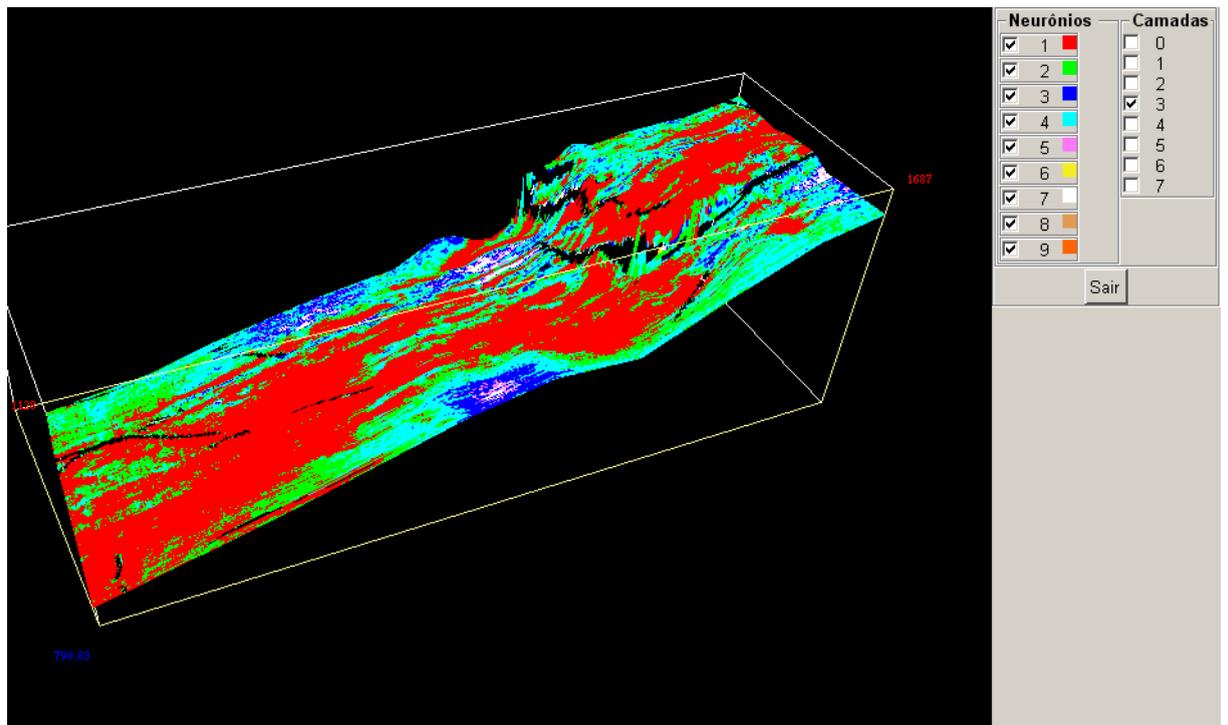


Figura 14 – Camada 3 – Algoritmo Kohonen

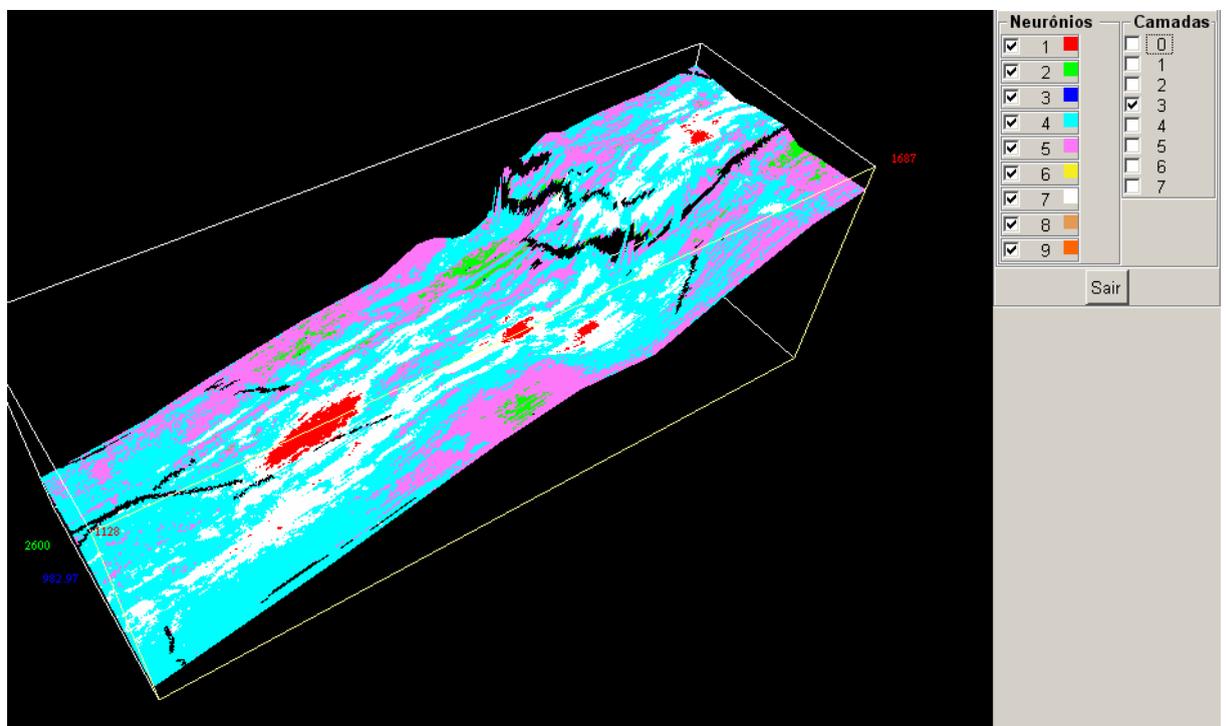


Figura 15 – Camada 3 – Algoritmo proposto

	Quantidade de elementos	Média	Desvio	Valor mínimo	Valor máximo
Grupo 1	127670	5815	176.16	4445	6002
Grupo 2	69484	6057	30.42	6002	6107
Grupo 3	68987	6318	31.07	6266	6374
Grupo 4	120973	6185	45.31	6107	6266
Grupo 5	21931	6392	11.00	6374	6412
Grupo 6	83750	6713	60.17	6612	6821
Grupo 7	97157	6509	58.18	6412	6612
Grupo 8	43286	6886	38.24	6821	6952
Grupo 9	139562	7115	146.48	6952	8097

Tabela 1 – Dados da classificação do algoritmo SOM de Kohonen

	Quantidade de elementos	Média	Desvio	Valor mínimo	Valor máximo
Grupo 1	5610	5255	185.22	4445	5460
Grupo 2	122796	6487	71.47	6370	6615
Grupo 3	84342	6718	60.84	6615	6827
Grupo 4	152241	5984	83.69	5809	6111
Grupo 5	184806	6234	73.76	6111	6370
Grupo 6	69129	6932	61.31	6827	7038
Grupo 7	42281	5695	86.87	5460	5809
Grupo 8	92034	7174	81.64	7038	7340
Grupo 9	19561	7436	80.29	7340	8097

Tabela 2 – Dados da classificação do algoritmo proposto

A visualização relativa às bordas (margens que circundam a estrutura) tem no algoritmo proposto uma melhor apresentação gráfica. A camada 1 dos dois processos é apresentada nas figuras 16 (SOM de Kohonen) e 17 (algoritmo proposto). Nos dois mapas é selecionado um trecho através de um retângulo para avaliação num detalhe maior nas figuras 18 e 19.

Comprando-se estes mapas observa-se que:

- É possível visualizar as formas das estruturas da figura 19 melhor do que as da figura 18;
- É possível visualizar em ambos os mapas as falhas de dados mostradas nas tarjas pretas.

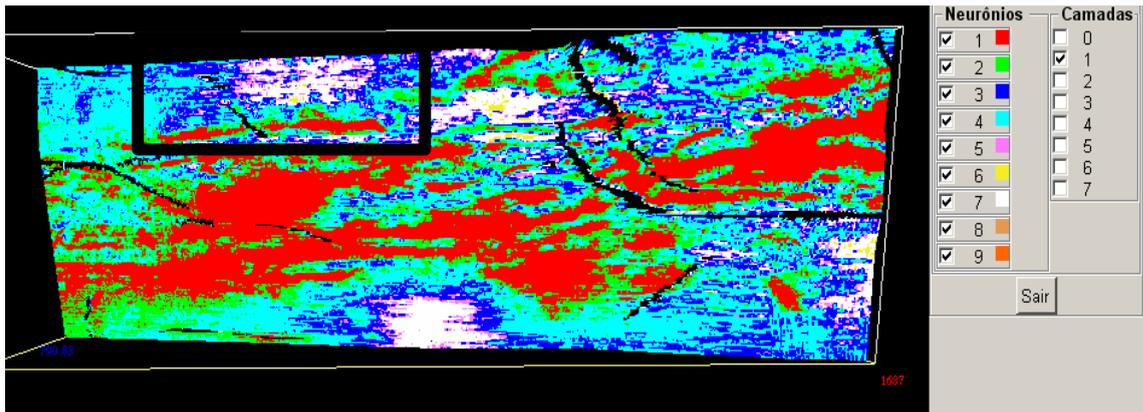


Figura 16 – Camada 1 – Algoritmo SOM Kohonen

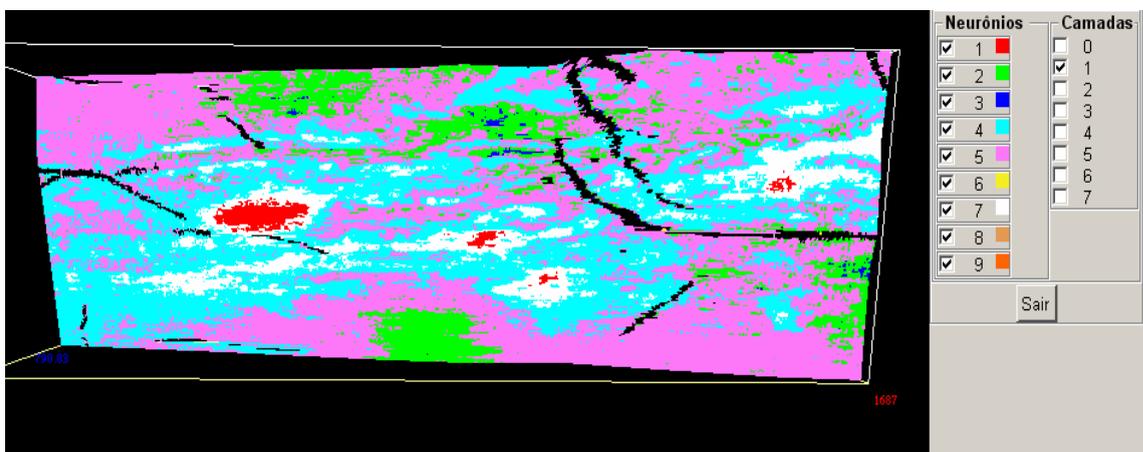


Figura 17 – Camada 1 – Algoritmo proposto

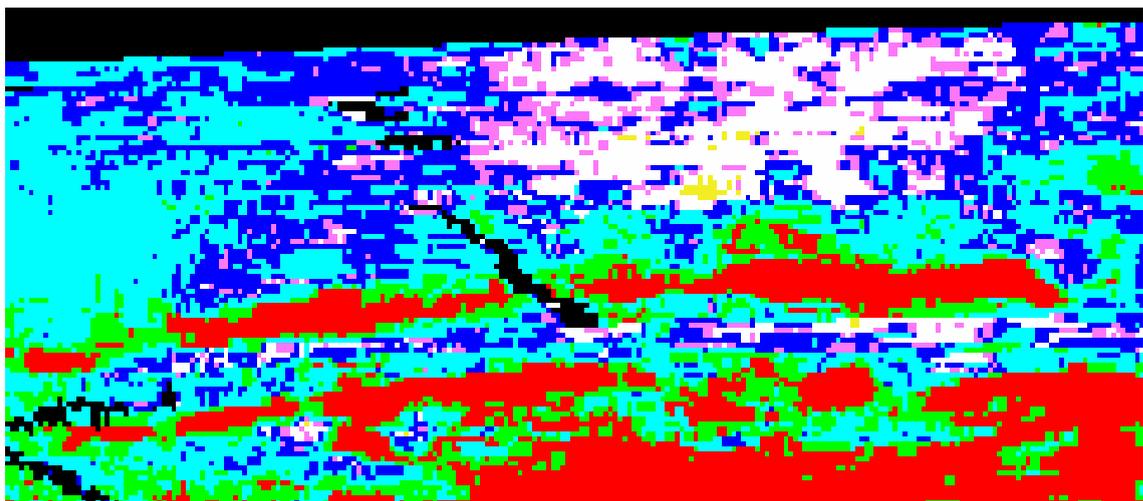


Figura 18 – Área selecionada – SOM Kohonen

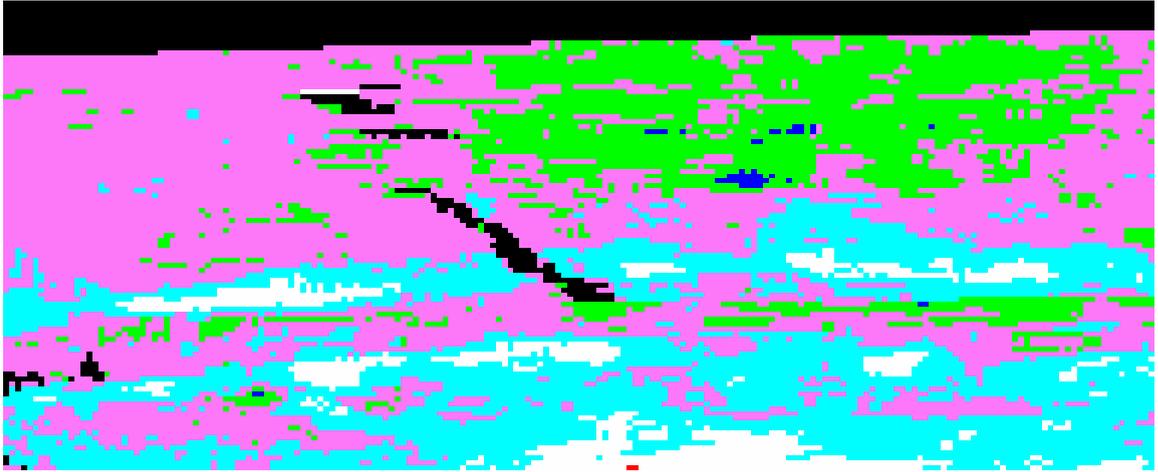


Figura 19 – Área selecionada – Algoritmo proposto

É possível afirmar, então, que as estruturas geológicas ficam melhores de serem visualizadas nos resultados do algoritmo proposto.

### 5.2.3. Identificação de Estruturas fantasmas

As estruturas fantasmas ocorrem por falha de captação de dados. Uma onda incide duplamente e reflete estruturas em tempos (eixo z) diferentes.

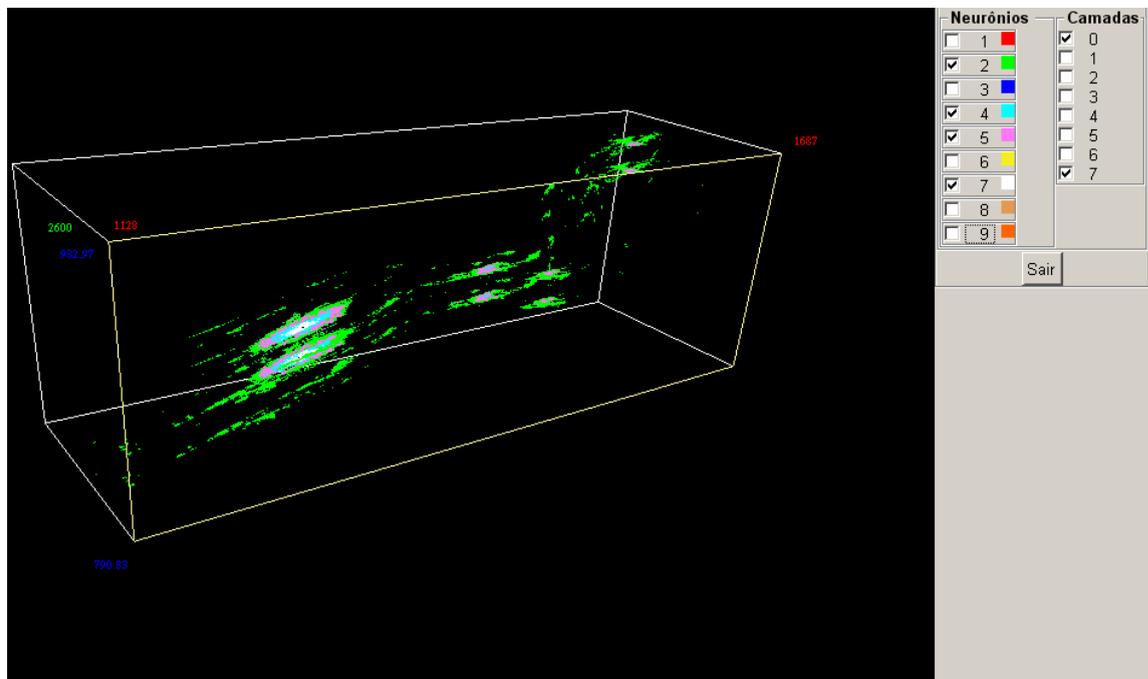


Figura 20 – Estruturas fantasmas

As estruturas fantasmas são um desafio para o tratamento dos dados sísmicos na medida em que, como é visto na figura, um dos conjuntos ocupa espaço de sinais que foram desprezados.

A maneira de identificar as estruturas fantasmas no método proposto é realizada avaliando-se a tabela de distribuições dos elementos por grupos e por camadas. A tabela 4 mostra esta distribuição

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
Camada 1	3	4484	13277	328	984	26861	160	43462	7041
Camada 2	1127	9251	258	34102	43367	7	8488	0	0
Camada 3	1456	4735	79	41039	37008	3	12280	0	0
Camada 4	1894	1935	19	48793	27257	1	16701	0	0
Camada 5	585	29845	3650	18539	40866	121	2987	7	0
Camada 6	348	41979	19646	7054	23468	2953	1044	107	1
Camada 7	190	27041	36463	2088	10972	16851	447	2533	15
Camada 8	7	3526	10950	298	884	22332	174	45925	12504
total	5610	122796	84342	152241	184806	69129	42281	92034	19561

Tabela 4 – Distribuição dos elementos por grupos e por camadas

O grupo 1 foi descartado pelo pequeno número de elementos. Os grupos 2, 4, 5 e 7 possuem valores próximos para as camadas 1 e 7 motivo pelo qual a verificação visual deve ser realizada.

#### 5.2.4. Visualização da Continuidade entre camadas

Os resultados do método proposto permitem a visualização da continuidade das estruturas propostas. A figura 21 mostra quatro blocos da estrutura correspondente ao grupo 3.

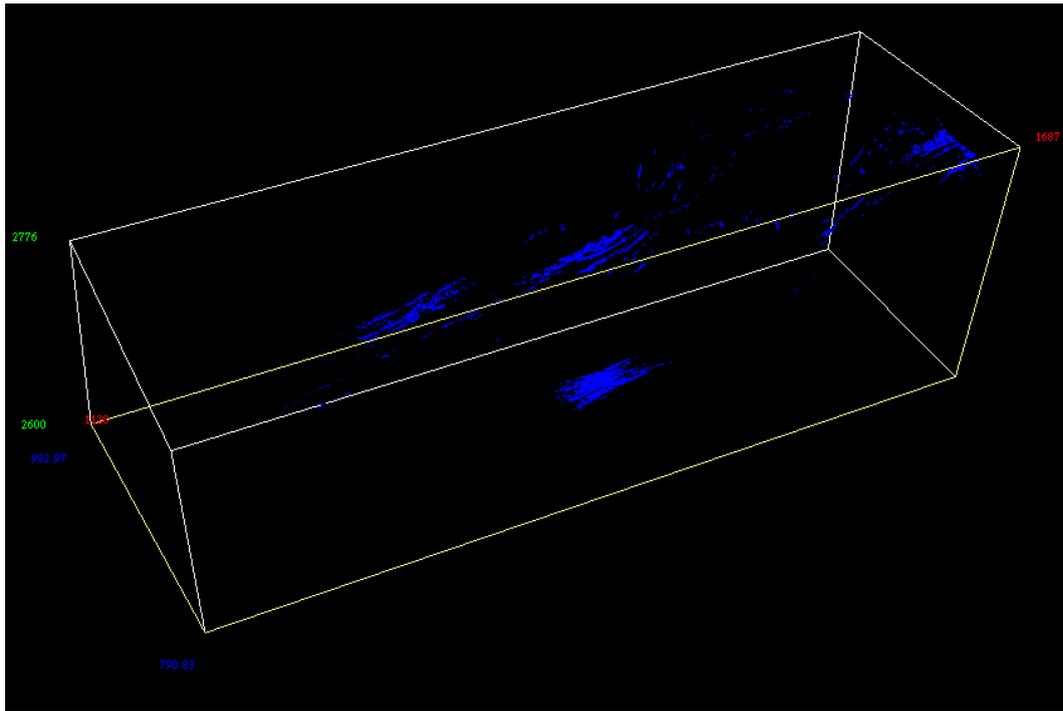


Figura 21a – Continuidade entre camadas – grupo 3 camada 4

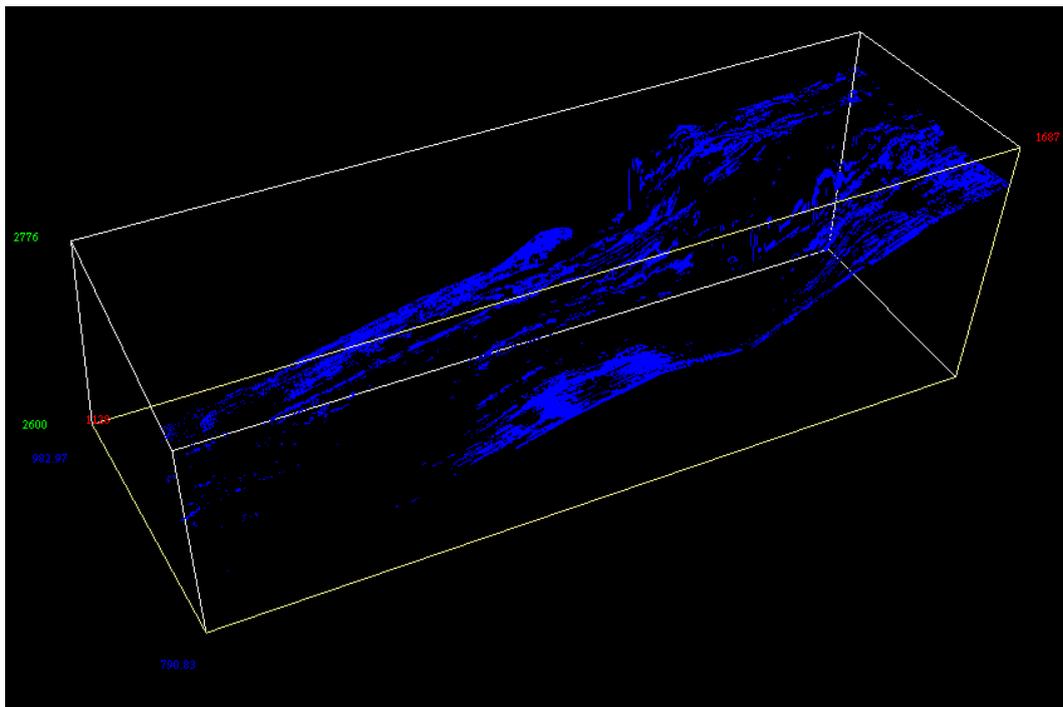


Figura 21b – Continuidade entre camadas – grupo 3 camada 5

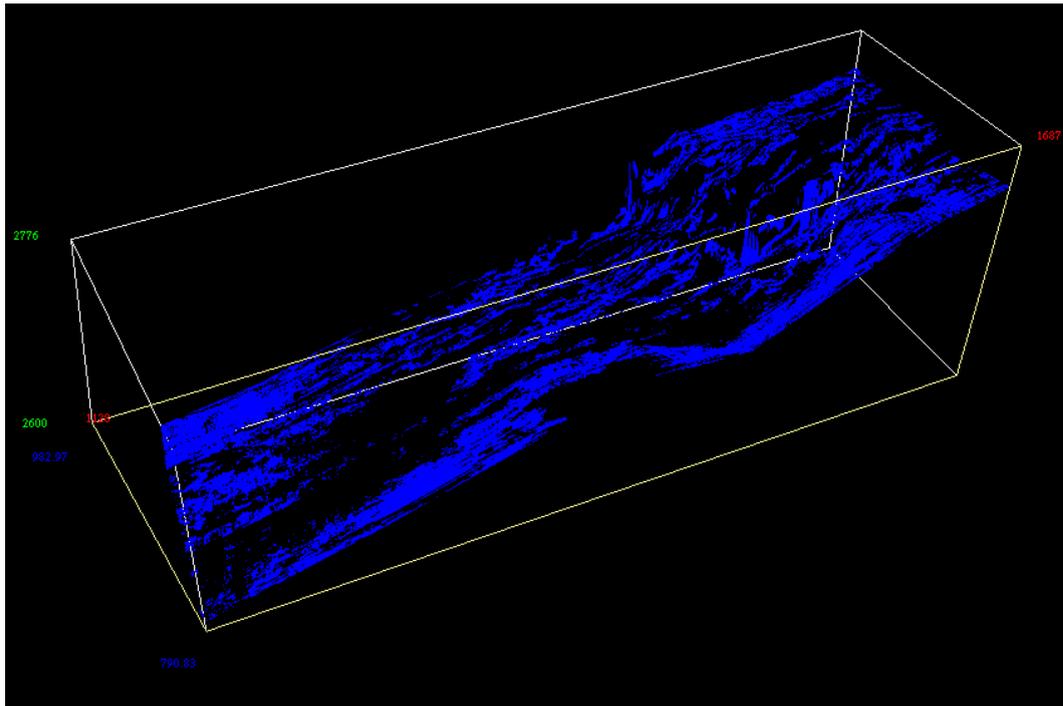


Figura 21c – Continuidade entre camadas – grupo 3 camada 6

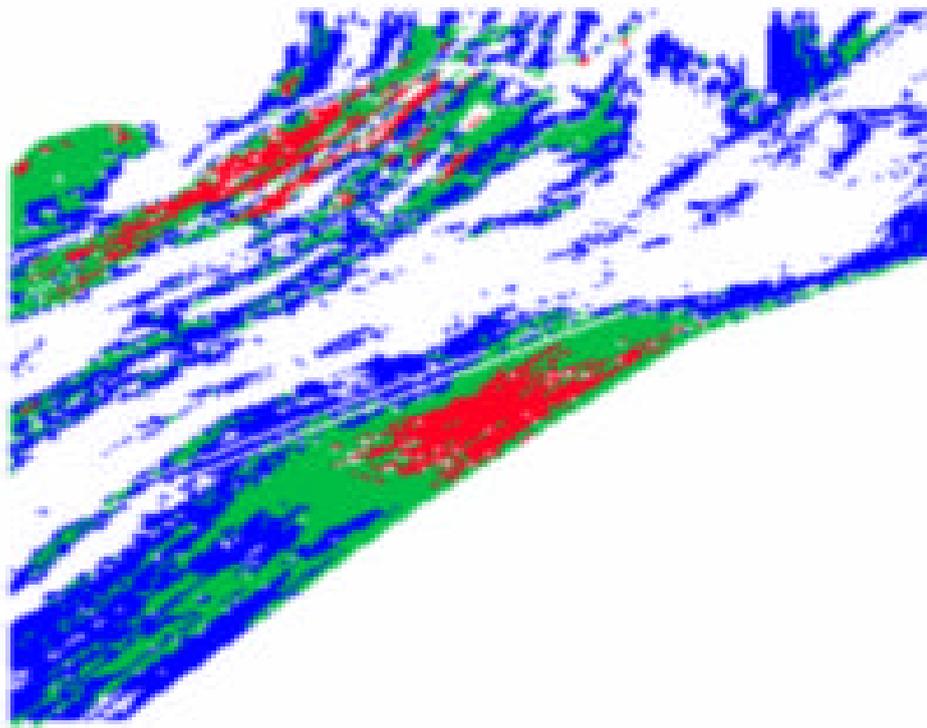


Figura 21d – Continuidade entre as camadas 4, 5 e 6

A figura 21d mostra, para área selecionada, uma superposição das três camadas com cores diferentes (vermelho para camada 4, verde para 5 e azul para a

6). Então é possível visualizar, mais facilmente, a continuidade da estrutura entre as camadas.