

**INTELIGÊNCIA ARTIFICIAL APLICADA AO RECONHECIMENTO DE  
PADRÕES LITOLÓGICOS**

Victor Ribeiro Carreira

Relatório apresentado ao Programa de Pós-graduação em Geofísica do Observatório Nacional, como parte dos requisitos necessários à obtenção do título de Doutor em Geofísica.

Orientador(a): Dr. Cosme F. Neto, Ponte

Rio de Janeiro  
Março de 2018

Resumo do Relatório apresentado ao Programa de Pós-Graduação em Geofísica do Observatório Nacional como parte dos requisitos necessários para a obtenção do título de Doutor em Geofísica.

## INTELIGÊNCIA ARTIFICIAL APLICADA AO RECONHECIMENTO DE PADRÕES LITOLÓGICOS

Victor Ribeiro Carreira

Março/2018

Este projeto propõe ...

# Sumário

<b>Lista de Figuras</b>	<b>iii</b>
<b>Lista de Tabelas</b>	<b>v</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Redes Neuronais Artificiais . . . . .	2
1.2 A Rede de Kohonen . . . . .	4
1.3 Redes com aprendizado não-supervisionado . . . . .	6
1.4 Medidas de Semelhança . . . . .	7
1.4.1 A métrica Euclideana . . . . .	7
1.4.2 A métrica de Mahalanobis . . . . .	8
1.4.3 Análise de agrupamento . . . . .	9
<b>2 Contexto Geológico</b>	<b>15</b>
<b>3 Método Proposto e Objetivo</b>	<b>17</b>
3.1 Objetivo . . . . .	20
<b>4 Dados de Perfilagem</b>	<b>21</b>
4.1 Modelo proposto para gerar os dados sintéticos . . . . .	24
4.2 Dado Real . . . . .	27
4.2.1 Dados reais e treinamento e classificação da rede . . . . .	28
<b>5 Resultados e Discussões</b>	<b>32</b>
5.1 Treinamento . . . . .	34
5.2 Identificação . . . . .	36
5.3 Dado Real: treinamento . . . . .	39
<b>6 Conclusões</b>	<b>46</b>
<b>7 Cronograma</b>	<b>48</b>
<b>Referências Bibliográficas</b>	<b>50</b>

# Listas de Figuras

1.1	Modelo esquemático de um neurônio de McCulloch-Pitts. Onde $x_1, x_2, \dots, x_n$ são os <i>inputs</i> , $w_1, w_2, \dots, w_n$ são os pesos, $h$ é o treino, $g(x)$ é a função de ativação, e $y$ é o <i>output</i> . . . . .	4
1.2	Neurônio e suas vizinhanças . . . . .	5
1.3	Homúnculo de Penfield. . . . .	6
1.4	Análise de agrupamento 1 . . . . .	10
1.5	Análise de agrupamento 2 . . . . .	11
1.6	Análise de agrupamento 3 . . . . .	13
2.1	Mapa geológico e de localização da área de estudo. . . . .	16
3.1	Fluxograma do programa de geração dos dados sintéticos . . . . .	17
3.2	Fluxograma do programa de treinamento da rede neuronal de kohonen. . . . .	18
3.3	Fluxograma do programa de identificação da rede. . . . .	19
4.1	Exemplo de um dado público de uma perfilagem de poço composta realizada pela Petrobras, na Bacia do Paraná. . . . .	22
4.2	Modelo Simplificado baseado em MOHRIAK <i>et al.</i> (2008). . . . .	25
4.3	Dado de perfilagem sintético, T1. Aonde a porcentagem de CE indica a mistura de conglomerado com embasamento . . . . .	26
4.4	Dado de perfilagem sintético, C1. . . . .	26
4.5	Dado de perfilagem sintético, C2. . . . .	27
4.6	Localização do total de poços de trabalho. . . . .	28
4.7	Localização dos poços escolhidos. . . . .	29
4.8	Poço 1BN0002SC e as respectivas propriedades físicas escolhidas para o teste. Em (b) tem-se a variação de litologia com a profundidade, (c) o potencial espontâneo e em (d) a resistividade lateral. . . . .	30
4.9	Poço 1BN0002SC e as respectivas propriedades físicas escolhidas para o teste. Em (b) tem-se a variação de litologia com a profundidade, (c) <i>Transient Time Integrator</i> e em (d) TOT. . . . .	31
5.1	Agrupamento de dados do poço T1. . . . .	32

5.2	Agrupamento de dados do poço C1. . . . .	33
5.3	Agrupamento de dados do poço C2. . . . .	34
5.4	Mapas auto-organizáveis e sua evolução temporal. . . . .	35
5.5	Teste de convergência da rede. . . . .	36
5.6	Dado de saída da rede para o poço de classificação C1. . . . .	37
5.7	Dado de saída da rede para o poço de classificação C2. . . . .	38
5.8	Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 40X40 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento. . . . .	40
5.9	Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 40X40 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento. . . . .	41
5.10	Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 40X40 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento. . . . .	42
5.11	Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 20X20 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento. . . . .	43
5.12	Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 20X20 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento. . . . .	44
5.13	Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 20X20 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento. . . . .	45

# **Lista de Tabelas**

1.1	Parâmetros dos teste analítico para comparação das métricas. . . . .	10
1.2	Parâmetros do segundo teste analítico. . . . .	12
1.3	Parâmetros dos teste analítico para comparação das métricas. . . . .	13
4.1	Compilação de Perfis usados na inferência de litologia. . . . .	23
4.2	Compilação de Perfis usados na inferência de porosidade, permeabilidade. . . . .	24
5.1	Tabela de referência para conversão do padrão numérico em litologia.	36
5.2	. . . . .	38
7.1	Cronograma das atividades previstas para o primeiro biênio. . . . .	48
7.2	Cronograma das atividades previstas para o segundo biênio. . . . .	49

# Capítulo 1

## Introdução

O ser humano vem usando a sua habilidade de reconhecimento de padrões desde muito antes do início do processo civilizatório. Grupos de humanos paleolíticos já faziam registro dos padrões migratórios de certos grupos de cervídeos. Durante a aurora da revolução neolítica, nossa capacidade de reconhecimento de padrões foi direcionada para a agricultura com a criação de monumentos que registraram a mudança das estações ao longo do ano.

O cérebro humano evoluiu espantosamente. E no que se refere a quantidade de informação processada, o cérebro possui enorme vantagem em relação a quantidade de informação processada por um computador (HALL *et al.*, 2014). Este não para de funcionar somente porque algumas células morrem. Um computador, por sua vez, não funciona quando há degradação da sua unidade central de processamento (MAO, 1996).

O campo do aprendizado de máquina aborda a criação de programas computacionais que automaticamente melhorem a si mesmos através da experiência (LEVY, 1997; MACKAY, 2005; MICHIE *et al.*, 1994).

As Redes Neuronais Artificiais (RNA) são inspiradas em modelos sensoriais do processamento de tarefas realizadas pelo cérebro (HAGAN *et al.*, 1996). Uma RNA, portanto pode ser criada através da aplicação de algoritmos matemáticos que imitem a tarefa realizada por um neurônio (NEDJAH *et al.*, 2016). Uma rede neuronal artificial possui semelhanças com a rede neuronal<sup>1</sup> natural presente no sistema nervoso central, neste o cômputo de informações realizado do cérebro é feito através de uma vasta quantidade de neurônios interconectados (FELDMAN *et al.*, 1988; POULTON, 2002). A comunicação entre essas células é realizada através de impulsos elétricos. Estes são transmitidos e recebidos por meio de sinapses nervosas entre

---

<sup>1</sup> Em muitas referências na área da inteligência artificial usa-se o termo neural ao invés de neuronal, contudo empregar o termo neuronal é um cuidado necessário e deve ser empregado no lugar do termo neural. Isso se deve ao fato de que os primeiros modelos matemáticos foram inspirados nas células e processos presentes no sistema nervoso central e não no sistema em toda a sua completude.

axônios e dendritos. As sinapses são estruturas elementares e uma unidade funcional localizada entre dois neurônios (KROGH, 2008).

Assim como as redes neuronais as medidas de similaridade são utilizadas como auxiliadores na predição da classe de um objeto. Ou, seja funciona como um algoritmo de aprendizado de máquina supervisionado que é basicamente aplicado em problemas de classificação(FREUND e MASON, 1999). A abordagem dos problemas de classificação sob a ótica de medidas de semelhança é um dos tópicos mais ativos dentro da área de aprendizado de máquina. O problema consiste em atribuir um rótulo a algum objeto baseado em um conjunto de atributos extraídos do mesmo. Para tal faz-se necessário, um conjunto de dados de treinamento com instâncias nas quais os rótulos dos objetos são conhecidos.

## 1.1 Redes Neuronais Artificiais

MCCULLOCH e PITTS (1943) redigem o trabalho pioneiro onde foi modelado um neurônio cuja resposta dependia do *input*<sup>2</sup> que provinha de outros neurônios e do peso utilizado. Já ROSENBLATT (1962) cria a teoria de convergência do *Perceptron* onde ele prova que modelos de neurônios possuem propriedades similares ao cérebro humano (KANAL, 2001). Neste sentido as rede neuronais artificiais podem realizar performances sofisticadas no reconhecimento de padrões, mesmo se alguns neurônios forem destruídos (LEVY, 1997). MINSKY e PAPERT (1969) demonstraram que *Perceptrons* somente resolvem uma classe muito limitada de problemas que podem ser linearizados.

Os primeiros artigos sobre redes neuronais em geofísica datam de 1989 e são focalizados basicamente na eficiência da RNA diante de dados distintos e como preparar esse dado para inserí-lo na RNA e posteriormente interpretá-lo. As redes neuronais artificiais foram usualmente treinadas com dados sintéticos e depois testados em dados reais. Contudo, hoje é comum usar dados reais para treinar a rede (ADIBI-FARD *et al.*, 2014). Embora, ambas as abordagens sejam aceitas. O foco a partir de 1995 até o presente relaciona-se a algumas aplicações específicas, tais como caracterização de reservatórios e na integração de dados associado a uma interpretação comprehensiva, ao contrário de uma aplicação isolada (POULTON, 2002).

No problema específicos de poços, um passo importante é a identificação de topo e base de camadas que podem ser associadas com mudanças das propriedades petrofísicas (SALJOOGHI e HEZARKHANI, 2014). Algoritmos baseados em derivadas nas curvas de log não identificam camadas muito finas, ou ruído (ZHANG *et al.*, 1999). CHAKRAVARTHY *et al.* (1999) consegue através do uso da função radial localizar os limites de camadas em alta definição em dados de log de indução

---

<sup>2</sup>Valor de entrada

(HDIL). Já BENAOUDA *et al.* (1999) consegue classificar tipos litológicos em poços parcialmente desmoronados através do uso da rede neuronal com propagação de erro e mudanças de classes a medida que prossegue a análise. CATÉ *et al.* (2017) levanta a questão da importância relativa das propriedades físicas, em dados de perfilagem de poços, para a tomada de decisão da rede neuronal.

O neurônio de MCCULLOCH e PITTS (1943) propõe um limite bimário para a criação de um modelo. Este neurônio artificial registra uma soma de pesos de  $n$  sinais de entrada,  $x_j$ ,  $j = 1, 2, 3, \dots, n$ , e fornece um *output*<sup>3</sup> de 1 caso esta soma esteja acima do limite  $u$ . Caso contrário o *output* é 0. Matematicamente essa relação pode ser descrita de acordo com a Eq. 1.1:

$$y = \theta \left( \sum_{j=1}^n w_j x_j - u \right) \quad (1.1)$$

Onde  $\theta$  é o passo dado na posição 0,  $w_j$  é chamada sinapse-peso associado a um *j*esimo *input*. A título de simplificação a função limite<sup>4</sup>  $u$  é considerada um outro peso  $w_0 = -u$  anexado a um neurônio com um *input* constante  $x_0 = 1$ . Pesos positivos correspondem a uma sinapse **excitatória**, enquanto pesos negativos correspondem a uma sinapse **inibitória**. Este modelo contém uma série de simplificações que não refletem o verdadeiro comportamento dos neurônios biológicos (MAO, 1996).

Derivações do neurônio de MCCULLOCH e PITTS (1943) na escolha das funções de ativação. Uma função largamente utilizada é a função sigmóide, que exibe uma suavização dos *outputs* a medida que o valor da função diminui (MAO, 1996; MISRA e SAHA, 2010). Essa função de ativação pode ser expressa de acordo com a Eq. 1.2:

$$g(x) = 1/(1 + e^{-\beta x}) \quad (1.2)$$

Onde  $\beta$  é o parâmetro de inclinação. A Fig. 1.1 ilustra a sequência lógica da operação de uma RNA para um neurônio simples de McCulloch-Pitts.

---

<sup>3</sup>Valor de saída

<sup>4</sup>Genericamente chamada de função de ativação

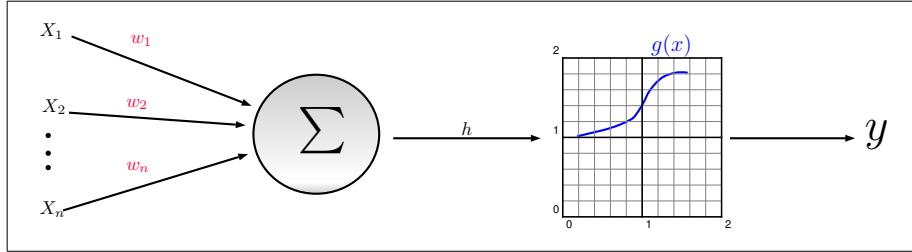


Figura 1.1: Modelo esquemático de um neurônio de McCulloch-Pitts. Onde  $x_1, x_2, \dots, x_n$  são os *inputs*,  $w_1, w_2, \dots, w_n$  são os pesos,  $h$  é o treino,  $g(x)$  é a função de ativação, e  $y$  é o *output*.

Mais de 50 tipos de redes neurais artificiais tem sido criadas até o ano de 2014 (SALJOOGHI e HEZARKHANI, 2014).

## 1.2 A Rede de Kohonen

Neste trabalho, foi utilizada a rede de kohonen. Esta rede neuronal tem como importante característica ser uma rede com aprendizado não-supervisionado, portanto o espaço solução de saída da rede não é conhecido.

A localização espacial de um neurônio da saída em um mapa topológico corresponde a um domínio ou característica particular do dado retirado do espaço de entrada. E estas entradas são mapeadas de forma ordenada, a exemplo dos mapas cito-arqueturais do córtex cerebral.

Neste processo de identificação de padrões a redundância torna-se impreverível, pois o neurônio da camada de saída que apresentar a maior resposta terá os seus pesos ajustados. Além disso, o peso dos neurônios vizinhos também serão ajustados em menor intensidade ao comparados com o neurônio vencedor.

Isto implica que os neurônios devem estar posicionados em um arranjo geométrico adequado. Esta teoria é baseada na suposição de que as células nervosas corticais estão organizadas anatomicamente em relação aos estímulos que recebem dos sensores aos quais estão ligadas (ARTERO, 2008).

Este modelo exige a definição de vizinhança entre neurônios de forma geométrica. Alguns arranjos são comumente utilizados, como por exemplo, os arranjos triangulares, hexagonal, retangulares, etc.

No caso de arranjos retangulares, diferentes vizinhanças de um neurônio  $N_{i,j}$  podem ser configuradas em quartetos, diagonais e octetos.

A Fig. 1.2 ilustra o arranjo retangular e as vizinhanças, em quartetos, diagonais e octetos, adotado neste trabalho.

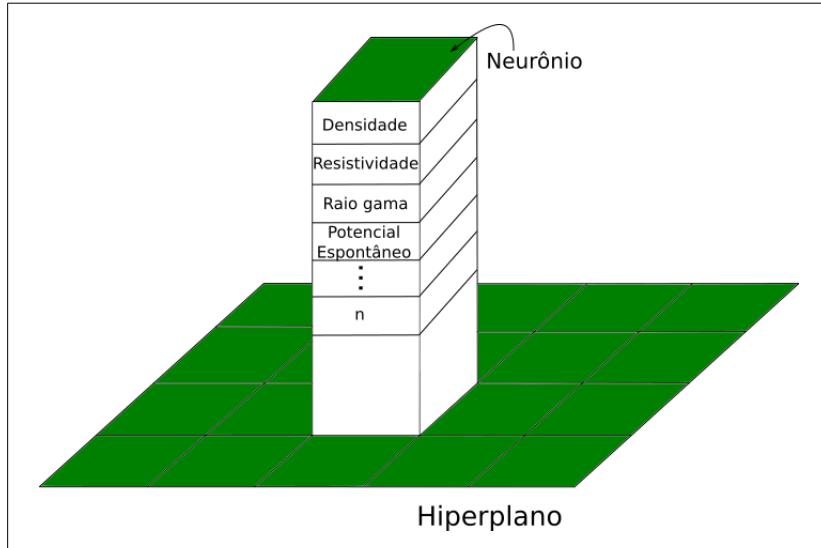


Figura 1.2: Neurônio e suas vizinhanças

O conceito de vizinhança representa uma competição pelo melhor aprendizado e o ajuste do vencedor e da sua vizinhança é um estímulo para que os neurônios ao redor do vencedor também melhorem.

Durante a etapa de treinamento é identificado o neurônio que tem os parâmetros de entrada mais parecidos com os valores dos pesos. Este procedimento é realizado via cálculo da distância euclidiana, Eq. 1.3, entre o parâmetro de entrada  $x(t)$  e o peso  $w_{i,j}$ .

$$d(t) = \sum_{i=1}^n [x(t) - w_{i,j}(t)]^2 \quad (1.3)$$

A etapa de treinamento da rede se dá por um ajuste de pesos entre os neurônios através do cálculo do menor valor de  $d(t)$  na iteração  $t$ , caracterizando assim o neurônio que passar por esse processo de *vencedor*. Esse procedimento ajusta da mesma forma os pesos do neurônio da vizinhança dentro. Os pesos são ajustados com uma fração da diferença entre os *inputs*  $x_i$  e os pesos  $w_i$ , vide Eq.1.4.

$$w_{i,j}(t+1) = w_{i,j}(t) + n(t)[x(t) - w_{i,j}] \quad (1.4)$$

Através deste ajuste continuado de pesos os elementos do conjunto de entrada são reorganizados de tal forma que as classes próximas sejam posicionados umas perto das outras. Isso gera um mapa bi-dimensional denominado na literatura de *mapa auto-organizável*. Este mapa é o análogo matemático mais fiel das áreas especializadas do córtex cerebral que são ilustradas pelo *Homúnculo de Penfield*, 1.3.

### 1.3 Redes com aprendizado não-supervisionado

Nesta categoria de RNA's são apenas inseridos os valores de *input* da rede. Os *output* são definidos pela própria rede que passa por um processo de treinamento não supervisionado. As redes que são submetidas a este tipo de treinamento são mais indicadas para tarefas aonde são exigidos agrupamento de dados (*clustering*). Neste processo uma classe deve ser atribuída aos registros da rede observando-se apenas o comportamento de seus atributos, no caso em particular deste trabalho tratam-se de propriedades geofísicas.

Uma rede com treinamento não supervisionado inspira-se no funcionamento do córtex cerebral. Neste modelo biológico, o organismo aprende a realizar alguma tarefa, por meio da identificação de padrões. Por exemplo, ao identificar uma música determinados padrões sonoros que compõe o conjunto harmonioso de notas precisam ser aprendidos antes de serem reconhecidos. Durante este processo, regiões específicas do cérebro vão sendo paulatinamente acionadas. Isto somente é possível, devido conexões específicas que são formadas entre os neurônios presentes no córtex, Fig. 1.3.

Os detalhes dos processos que regulam o córtex ainda não foram totalmente elucidados, contudo é seguro assumir que a primeira representação dos fenômenos de aprendizagem podem ser representados por uma superfície topológica ou mapa auto-organizado.

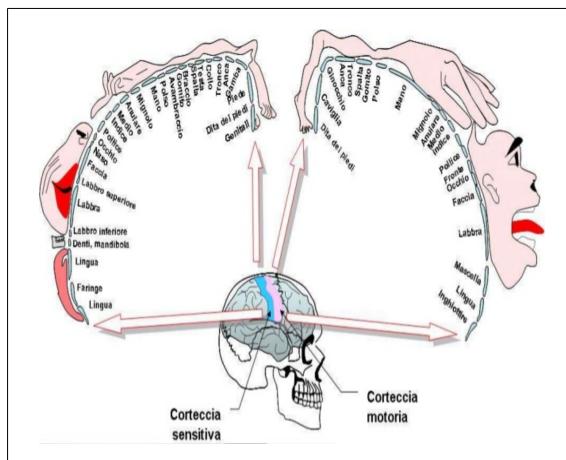


Figura 1.3: Homúnculo de Penfield.

Um cérebro que sofreu uma comoção grave perde a capacidade de acessar determinadas zonas do homúnculo responsáveis por atividades específicas. Contudo o cérebro tem a capacidade de destinar outras regiões para o controle destas ações que foram previamente perdidas.

Além de casos graves como um acidente o cérebro também perde a capacidade de aprendizado com o tempo. Em humanos, a capacidade de aprendizado vai da

pequena infância até a puberdade. Após este período, o cérebro passar a reter o que fora aprendido. Sendo assim o aprendizado é uma função que depende, entre outras coisas, do tempo.

## 1.4 Medidas de Semelhança

Dado um conjunto de dados numéricos, a **métrica do dado** é qualquer leitura em uma dada escala de intervalo que infere o grau de diferença entre dois objetos. Dados que são caracterizados como dados **não-métricos**, são aqueles conjuntos de dados que podem ser coletados em um formato *binário* (0/1), *ordinário* números que expressam uma posição somente (índice), ou *escala nominal* que são conjuntos de dados não ordenados (MICHEL e DEZA, 2016).

Na análise geométrica do dado se refere aos aspectos geométricos da imagem, análise de padrões ou forma, que trata um conjunto arbitrário de dados como uma nuvem de pontos no espaço  $\Re^n$ . O dado passa a ser organizado em uma base de dados indexadas em um espaço métrico<sup>5</sup>.

Na análise de agrupamentos (classificação, taxonomia, reconhecimento de padrões) consiste em dividir o dado  $A$  em um conjunto menor de grupos. Por exemplo, um grupo de dados que estão próximos em respeito a um determinado critério como propriedades físicas em rochas.

Neste trabalho são estudadas duas medidas de semelhança especiais. A primeira delas é a métrica de *Euclides* que leva em consideração o cálculo de centroides para avaliar a distância entre dois agrupamentos. A segunda é a métrica de *Mahalanobis* que leva em consideração a forma do agrupamento a ser analisado.

### 1.4.1 A métrica Euclideana

Dado dois conjuntos de agrupamentos distintos,  $\theta(\bar{x}_i, \bar{y}_i)$  e  $\beta(\bar{x}_j, \bar{y}_j)$ , os seus vetores de coordenadas cartesianas podem ser definidos como

$$\mathbf{x}^i = \begin{bmatrix} x_1^i \\ x_2^i \\ x_3^i \\ \vdots \\ x_m^i \end{bmatrix} \quad (1.5) \qquad \mathbf{y}^i = \begin{bmatrix} y_1^i \\ y_2^i \\ y_3^i \\ \vdots \\ y_m^i \end{bmatrix} \quad (1.6)$$

para um agrupamento  $\theta$ , e

---

<sup>5</sup>Este processo é conhecido como indexação métrica.

$$\mathbf{x}^j = \begin{bmatrix} x_1^j \\ x_2^j \\ \vdots \\ x_n^j \end{bmatrix} \quad (1.7) \qquad \mathbf{y}^j = \begin{bmatrix} y_1^j \\ y_2^j \\ \vdots \\ y_n^j \end{bmatrix} \quad (1.8)$$

para outro agrupamento  $\beta$ . A métrica euclideana, na sua forma mais genérica, pode ser definida como uma distância ponderada, de acordo com a Eq. 1.9.

$$[(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)^T \mathbf{A} (\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j)]^{1/2} \quad (1.9)$$

Onde  $\mathbf{A}$  é uma matriz não-singular e simétrica de dimensão  $m \times n$  e com  $m = n$ ,  $\bar{\mathbf{x}}_i$  é a média do vetor de coordenadas cartesianas do agrupamento  $\theta$  com dimensão  $m$ ,  $\bar{\mathbf{y}}_i$  é a média do segundo vetor de coordenadas do agrupamento  $\theta$  com dimensão  $m$ ,  $\bar{\mathbf{x}}_j$  é a média do vetor de coordenadas do agrupamento  $\beta$  com dimensão  $n$ ,  $\bar{\mathbf{y}}_j$  é a média do segundo vetor de coordenadas cartesianas do agrupamento  $\beta$  também com dimensão  $n$ .

Nos casos em que  $\mathbf{A} = \mathbf{I}$ , onde  $\mathbf{I}$  é a matriz identidade obtém-se a métrica euclideana não-ponderada, que é o caso estudado neste trabalho.

A métrica euclideana é de fácil compreensão. Contudo sua utilização apenas apresenta bons resultados, quando todas as classes possuam a mesma variância. Além disso, os elementos dos vetores não devem possuir nenhum tipo de correlação entre si, em resumo, estes devem ser circulares. E no espaço de propriedades, onde os elementos dos vetores se distribuem, isso raramente ocorre.

### 1.4.2 A métrica de Mahalanobis

A distância de Mahalanobis é um caso particular da métrica euclideana onde, no lugar da matriz  $\mathbf{A}$ , utiliza-se a matriz de covariância agrupada inversa  $\mathbf{S}$ . Esta métrica mede a separação entre dois grupos de objetos levando-se em consideração o formato da distribuição dos dados. Suponhamos que nós tenhamos dois grupos de objetos com médias  $\bar{\mathbf{x}}_i$  e  $\bar{\mathbf{x}}_j$ , a distância de Mahalanobis é dado pelo seguinte enunciado, Eq. 1.10:

$$[(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)^T \mathbf{S}^{-1} (\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j)]^{1/2} \quad (1.10)$$

Os dados dos dois grupos devem ter o mesmo número de variáveis (o mesmo número de colunas), mas não necessariamente o mesmo número de dados (cada grupo pode possuir diferentes número de linhas). A matriz covariância para o grupo  $i$  é calculada usando uma matriz de dados centralizada  $\hat{\mathbf{X}}$ .

$$\mathbf{C}_i = \frac{1}{n} \hat{\mathbf{X}}_i^T \hat{\mathbf{X}}_i \quad (1.11)$$

Onde  $n = \sum_i n_i$ , ou seja a soma de todos os dados de todos os grupos.

A matriz de covariância agrupada  $\mathbf{S}$  (*Pooled Covariance Matrix*) dos dois agrupamentos ( $\theta, \beta$ ) é computada como a média ponderada das matrizes de covariância:

$$\mathbf{S}(\theta, \beta) = \frac{1}{n} \sum_{i=1}^g n_i C_i \quad (1.12)$$

Onde  $\theta$  e  $\beta$  representam os dois agrupamentos e  $g$  é o número de agrupamentos.

### 1.4.3 Análise de agrupamento

Ao se criar uma análise de agrupamentos tem-se em vista a caracterização do quão semelhante, ou não, dois ou mais conjuntos de dados podem ser. Para se determinar em qual situação cada métrica se comporta melhor, foi gerado um teste analítico com vistas a se comparar tanto a distância quanto a forma da distribuição dos dados influem no valor absoluto final das duas métricas.

No primeiro teste, gerou-se distribuições randômicas circulares de três agrupamentos com seus centros igualmente espaçados como apresentado na Fig. 1.4. Nota-se que os clusters encontram-se a 20 unidades de distância entre seus respectivos centros.

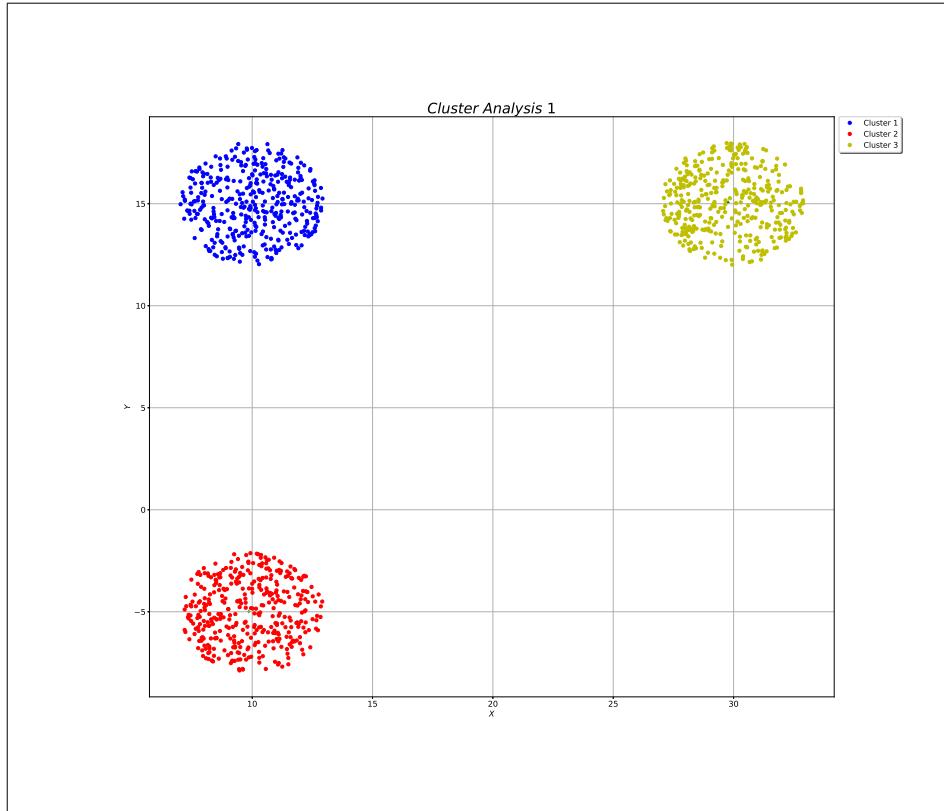


Figura 1.4: Análise de agrupamento 1

O primeiro caso estudado apresentam três circunferências compostas por diferentes conjuntos de dados. Os parâmetros utilizados para criar a primeira situação de estudo são apresentadas na Tab. 1.1.

Tabela 1.1: Parâmetros dos teste analítico para comparação das métricas.

Parâmetros	Cluster 1	Cluster 2	Cluster 3
Raio máximo	0.300E+01	0.300E+01	0.300E+01
Raio mínimo	0.000E+00	0.000E+00	0.000E+00
Ângulo máximo	0.157E+02	0.157E+02	0.157E+02
Ângulo mínimo	0.000E+00	0.000E+00	0.000E+00
Semente	5	5	5
Número de pontos	404	383	397

Para os casos em que a distribuição, no espaço de propriedades, apresenta simetria em ambos os eixos as métricas apresentam uma pequena diferença entre si. A distância de Euclides medida entre os agrupamentos I e II foi de 20,0 unidades de medida, enquanto que a mesma distância medida entre os clusters I-III foi de 19,8 unidades de medida. Tal diferença se deve a imprecisão associada ao sorteio randômico e o deslocamento dos centros das distribuições.

A distância de Mahalanobis leva em consideração a forma das distribuições de

dados ao longo do espaço de propriedades. O resultado desta métrica para os agrupamentos I-II foi de 13,8, enquanto que a mesma métrica para os agrupamentos I-III foi de 12,9. Mostrando que os resultados se afastam um pouco mais.

Distâncias	
Euclideana (I-II)	0.200E+02
Mahalanobeana (I-II)	0.138E+02
Euclideana (I-III)	0.197E+02
Mahalanobeana (I-III)	0.129E+02

O segundo teste apresenta uma leve deformação no agrupamento I (em azul), Fig. 1.5. Um pequeno aumento do eixo horizontal muda a distribuição para uma forma levemente elipsoidal de dados no espaço. Os demais clusters continuam com uma distribuição circular.

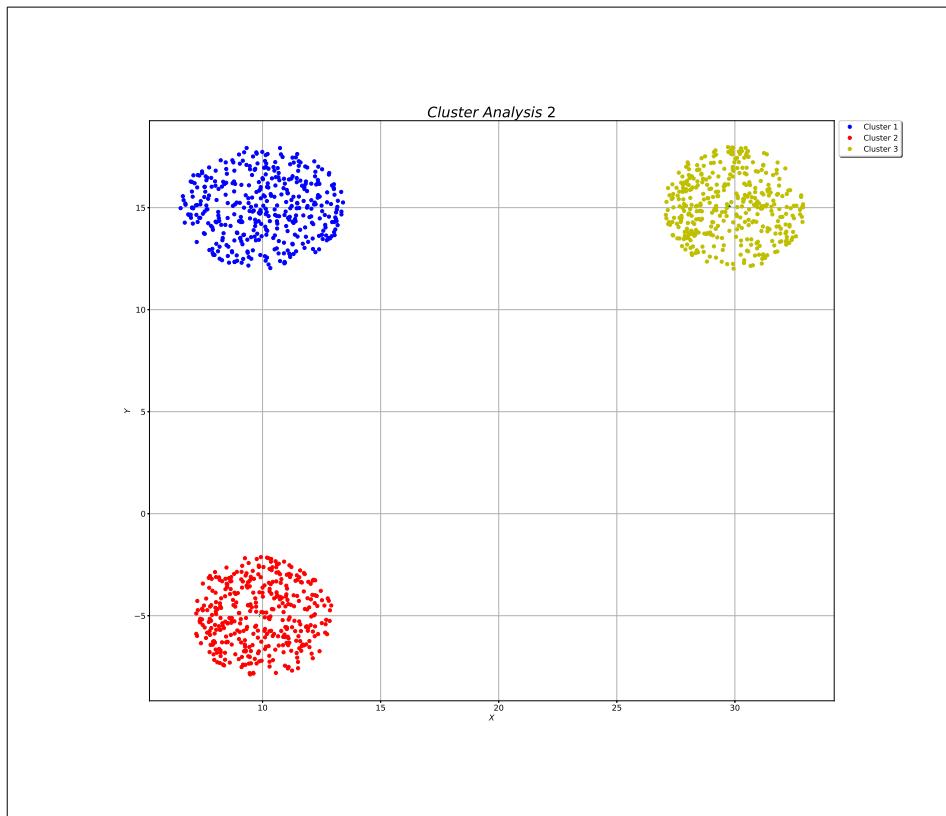


Figura 1.5: Análise de agrupamento 2

A Tab. 1.2 mostram os parâmetros utilizados para a criação de cada um dos clusters. Salienta-se a principal mudança foi um leve aumento do eixo horizontal da ordem de 0,5 unidades de medida. Os demais parâmetros tais como posição e número de dados dos clusters permaneceram inalterados.

Tabela 1.2: Parâmetros do segundo teste analítico.

Parâmetros	Cluster 1	Cluster 2	Cluster 3
Raio máximo	—	0.300E+01	0.300E+01
Raio mínimo	—	0.000E+00	0.000E+00
Ângulo máximo	0.157E+02	0.157E+02	0.157E+02
Ângulo mínimo	0.000E+00	0.000E+00	0.000E+00
Eixo maior	0.350E+01	—	—
Eixo menor	0.300E+01	—	—
Semente	9	9	9
Número de pontos	405	383	396

No segundo caso, a distância de Euclides permaneceu inalterada para ambos os casos, tanto na distância entre os agrupamentos I-II e I-III. Isso se deve ao fato de os centroides permanecerem inalterados para as análises de agrupamentos 1 e 2, observando-se um valor de 20,0 e 19,7 unidades de distância. A distância de Mahalanobis sofreu o maior impacto no seu valor absoluto, na ordem de 2 unidades de medida entre os agrupamentos I-III. De maneira geral os padrões de comportamento das métricas permaneceram inalterados, com exceção do caso aonde a Mahalanobis foi de 11,8 unidades de medida demonstrando a sensibilidade esperada para deformações em grupos de dados. Enquanto que o valor encontrado para o agrupamento I-II foi de 13,8, muito semelhante ao do teste anterior.

Distâncias	
Euclideana (I-II)	0.200E+02
Mahalanobeana (I-II)	0.138E+02
Euclideana (I-III)	0.197E+02
Mahalanobeana (I-III)	0.118E+02

O terceiro teste apresenta o agrupamento 1 com um formato elipsoidal oblato se aproximando mais do agrupamento 3. Este comportamento é retratado na Fig. 1.6. Os demais conjuntos de agrupamentos permaneceram com a mesma forma dos testes anteriores.

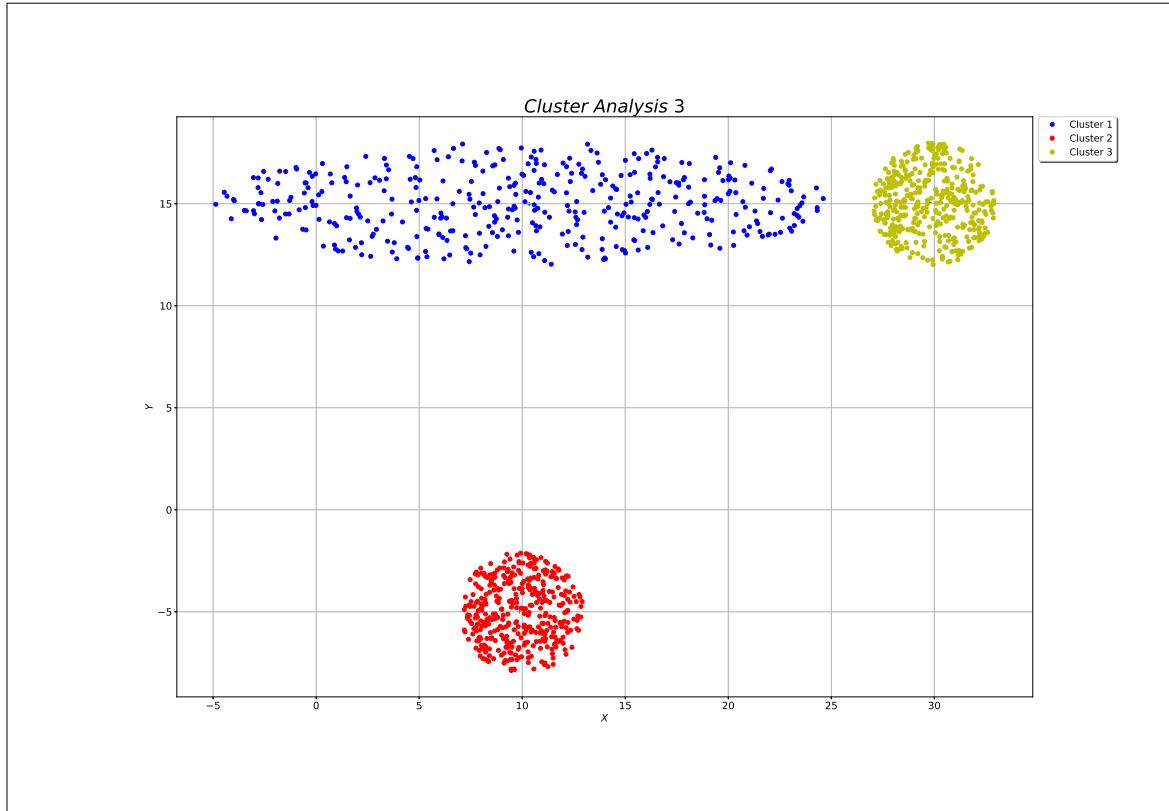


Figura 1.6: Análise de agrupamento 3

A deformação explícita no agrupamento 1 se deu por conta do aumento de 11,5 unidades de medida do eixo horizontal. Essa perturbação na forma da elipse em contraste com os demais clusters é descrita na Tab. 1.3.

Tabela 1.3: Parâmetros dos teste analítico para comparação das métricas.

Parâmetros	Cluster 1	Cluster 2	Cluster 3
Raio máximo	—	0.300E+01	0.300E+01
Raio mínimo	—	0.000E+00	0.000E+00
Ângulo máximo	0.157E+02	0.157E+02	0.157E+02
Ângulo mínimo	0.000E+00	0.000E+00	0.000E+00
Eixo maior	0.150E+02	—	—
Eixo menor	0.300E+01	—	—
Semente	17	17	17
Número de pontos	405	384	395

Distâncias	
Euclideana (I-II)	0.200E+02
Mahalanobeana (I-II)	0.138E+02
Euclideana (I-III)	0.194E+02
Mahalanobeana (I-III)	0.352E+01

A distância de Euclides permaneceu inalterada dos três casos estudados entre os três agrupamentos. Apresentando o valor de 2,0 e 19,4 para as distâncias entre os agrupamentos I-II e I-III respectivamente. Contudo, o cômputo da distância de Mahalanobis apresentou uma grande diferença se comparado as análises anteriores. Apresentando o valor de 13,8 unidades de medida entre os agrupamentos I-II e 3,5 unidades de medida entre os agrupamentos I-III.

Esta pequena análise demonstra que o cômputo da distância de Mahalanobis está intimamente relacionada com a forma da distribuição dos dados dentro do espaço de propriedades quando comparada a distância Euclideana. Isto se deve ao cálculo da matriz de covariância para cada classe de dados 1.10, que permite levar em conta a forma do agrupamento em questão.

# Capítulo 2

## Contexto Geológico

A Bacia do Paraná desenvolveu-se sobre uma área de escudo do continente Gondwana Sul e é composta por uma série de núcleos cratônicos, rodeados por vários cinturões móveis e cobertos por bacias molássicas, que foram desenvolvidas durante o ciclo termo-tectônico Brasiliano que se estendeu desde o neoproterozóico até o Ordoviciano. A deformação decorrente deste ciclo teve início entre 700 Ma e 650 Ma, sendo que a maior parte das intrusões de granitos que podemos observar na Bacia, situou-se dentro do limite entre o Proterozóico e o Paleozóico (cerca de 570 Ma) com resfriamento durante o Cambro-Ordoviciano entre 500 – 450 Ma (HAWKESWORTH *et al.*, 2000; ZALAN e WOLF, 1987).

O embasamento que circunda a Bacia do Paraná é dividido em: margem Leste/Sudeste, representado pelas faixas Dom Feliciano e Ribeira ,de idade Brasiliana e de direção NE-SW, separados por um núcleo cratônico designado Rio de La Plata/ Luiz Alves; margem Norte/Nordeste, representada pela faixa Uruaçu, de idade mesoproterozóica, de direção NW e por dois maciços arqueanos (Guaxupé e Goiás) remobilizados durante o ciclo Brasiliano; margem Oeste/Noroeste representada pela faixa de dobramentos Paraguai/Araguaia, também do ciclo Brasiliano, que delimita o extremo da borda Noroeste da Bacia (BORGHI, 2002; HAWKESWORTH *et al.*, 2000).

Dentre os principais grupos de estruturas, nota-se três grupos de lineamentos de direções preferenciais NW-SE, E-W e NE-SW, representando cada um evento termo-tectônico distinto. O conjunto de lineamentos NW-SE são os mais antigos e estão relacionados ao evento termo-tectônico do Transamazônico, e, as zonas de falhas geológicas associadas a este evento foram reativadas durante o rifteamento do Atlântico Sul, no Cretáceo. Os lineamentos E-W, tiveram início a partir do Triássico e são paralelos às zonas de fratura oceânica, sugerindo uma ligação com o desenvolvimento do Atlântico Sul. Os lineamentos NE-SW são derivados do evento tremo-tectônico Brasiliano e de seus cinturões móveis associados. Este último conjunto de lineamentos é isento de diques de basalto (MILANI e ZALAN, 1999).

O registro estratigráfico da Bacia do Paraná é formado por pacote sedimentar e magmático de espessura máxima em torno de 7000 m, que coincide geograficamente com o depocentro estrutural da sinéclise e com a calha do rio paraná (MILANI e RAMOS, 1998b). O registro estratigráfico da Bacia do Paraná é dividido em seis unidades de ampla escala ou supersequências (VAIL *et al.*, 1977) na forma de pacotes rochosos com intervalos temporais de algumas dezenas de milhões de anos de duração e envelopados por superfícies de discordância de caráter inter-regional: Rio Ivaí (Ordoviciano-Siluriano), Paraná (Devoniano), Gondwana I (Carbonífero-Eotriássico), Gondwana II (Meso a Neotriássico), Gondwana III (Neojurássico-Eocretáceo) e Bauru (Neocretáceo). As três primeiras supersequências são representadas por sucessões sedimentares que definem ciclos transgressivos e regressivos ligados às oscilações do nível relativo do mar, durante o Paleozóico, ao passo que as demais correspondem a pacotes de sedimentos continentais com rochas ígneas associadas. As unidades formais da litoestratigrafia, quais sejam os grupos, formações e membros comumente utilizados na descrição do arranjo espacial dos estratos da bacia, inserem-se como elementos particularizados neste arcabouço aloestratigráfico de escala regional (MILANI *et al.*, 2007).

O mapa geológico (Fig. 2.1) apresenta a extensão e os limites da Bacia do Paraná (BIZZI *et al.*, 2003).

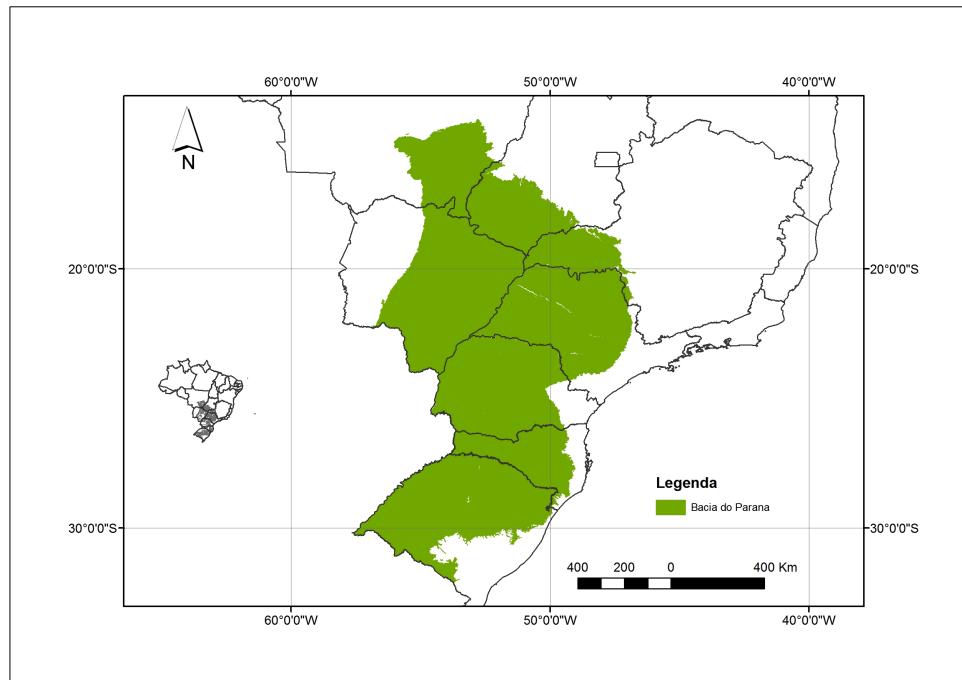


Figura 2.1: Mapa geológico e de localização da área de estudo.

# Capítulo 3

## Método Proposto e Objetivo

A parte operacional da metodologia divide-se em três etapas: 1- Geração de dados sintéticos, 2- Treinamento e 3- Identificação. Cada uma destas etapas será realizada por um programa computacional específico, estes programas vão funcionar de forma independente.

O primeiro programa tem por objetivo gerar dados sintéticos que devem simular os resultados obtidos num levantamento de um perfil composto.

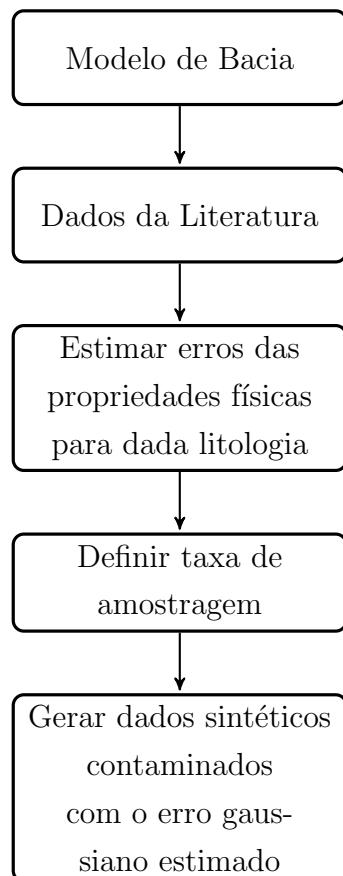


Figura 3.1: Fluxograma do programa de geração dos dados sintéticos

O programa da etapa de Treinamento será alimentado com dados de perfilagens cujas respectivas fácies litológicas são conhecidas (initialmente serão usados dados sintéticos e posteriormente dados reais). Este programa vai gerar um arquivo com os dados do treinamento, que será usado pelo programa de Operação. Esta é a fase de aprendizagem da rede.

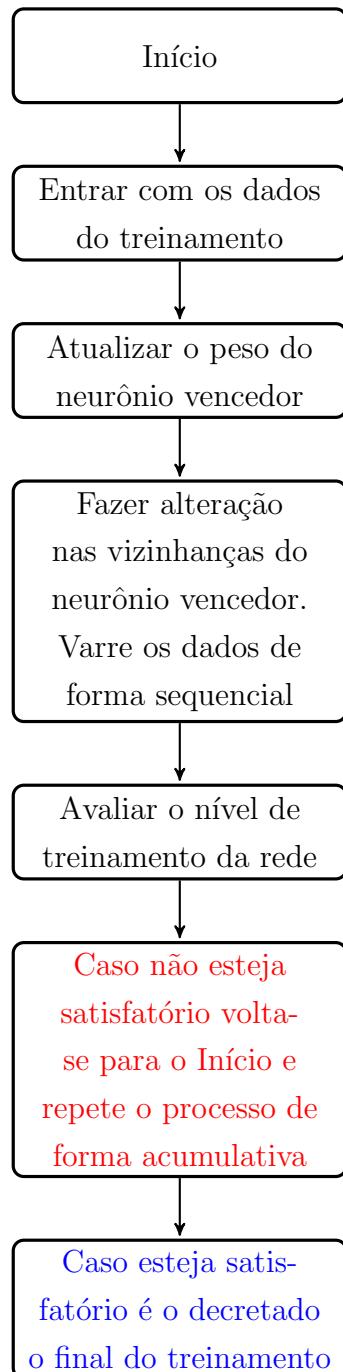


Figura 3.2: Fluxograma do programa de treinamento da rede neuronal de Kohonen.

O programa da etapa de identificação vai fazer a classificação, de forma autônoma, das facies litológicas em poços a partir dos dados de perfilagem em poços nos quais

a litologia é desconhecida. A aprendizagem da rede deve ocorrer de forma continuada, quanto mais informação temos sobre situações nas quais a litologia é conhecida mais bem preparada estará a rede em termos de aprendizagem. Este conceito de aprendizagem é acumulativo e isso ocorrerá através da atualização do arquivo com os dados de treinamento.

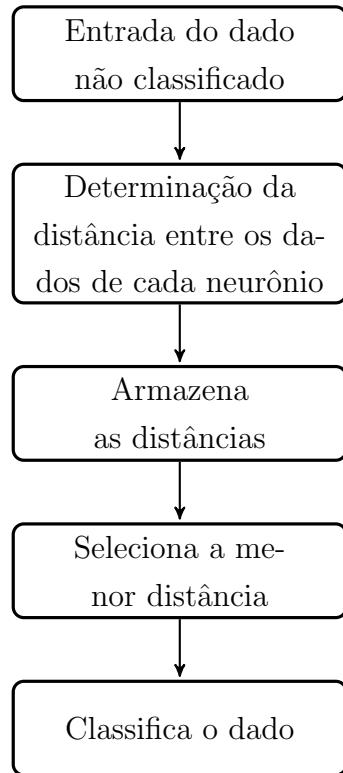


Figura 3.3: Fluxograma do programa de identificação da rede.

Durante a elaboração dos programas será necessário testar a sua eficiência. Estes testes serão realizados através de dados sintéticos que serão gerados por um terceiro programa, gerador de dados sintéticos. Este programa será alimentados com informações da literatura. Após os testes com dados sintéticos a metodologia será validada com dados reais, posteriormente depois de cumpridas todas estas etapas, a metodologia estará pronta para ser utilizada em situações reais.

É importante salientar que neste método o conhecimento do funcional geofísico, que rege a relação entre litologia e as propriedades físicas das rochas, não é necessário durante o processo. O conceito de inteligência artificial que será utilizado prescinde do funcional geofísico, o aprendizado é feito através da identificação de padrões recorrentes, sendo este ponto positivo na metodologia, pois o funcional geofísico, que por vezes é desconhecido ou de alta complexidade, exige uma modelagem matemática custosa.

O ponto negativo é a necessidade de se ter muitos dados já analisados em situações conhecidas e variadas para a realização da etapa de treinamento. A etapa

de treinamento tem um custo computacional alto.

### 3.1 Objetivo

O principal objetivo deste projeto é desenvolver um programa computacional do tipo “ machine learning ”, que será implementado na forma de uma Rede Neuronal Artificial (RNA) dentro do contexto da inteligência artificial. Este programa deve ter a capacidade de identificar, de forma autônoma, fácies litológicas a partir de dados de perfilagem de poços sem a necessidade do uso de um funcional geofísico.

É importante salientar que a metodologia que será desenvolvida neste projeto tem aplicação direta tanto na indústria de exploração mineral, quanto na de água, e na de petróleo e gás.

# Capítulo 4

## Dados de Perfilagem

As RNAs são capazes de reconhecer padrões (KONATÉ *et al.*, 2014; KUMAR *et al.*, 2015). E padrões muitas vezes são recorrentes no tocante a geologia (VAIL *et al.*, 1977).

Ciclos de deposição de siltes e argilas e areias muitas vezes são controlados pelas variações constantes das estações do ano (CRISTINA LOPES QUINTAS *et al.*, 1999; MILANI e RAMOS, 1998a; MILANI *et al.*, 2000) . Esse registro litológico se faz presente em dados de poços em todo o mundo (SCHERER e LAVINA, 2006).

Em uma perfilagem de poço composta são realizadas diversas medidas de propriedades físicas que ao serem analisadas, em conjunto, tornam possível ao geólogo identificar mudanças litológicas e consequentemente topos e bases de camadas de interesse (ARTUR e SOARES, 2008; FRANCA, ALMERIO & POTTER, 1991; ZALAN, 2007).

A Fig. 4.1 ilustra a disposição de um perfil de poço associado com topos e bases de rochas.

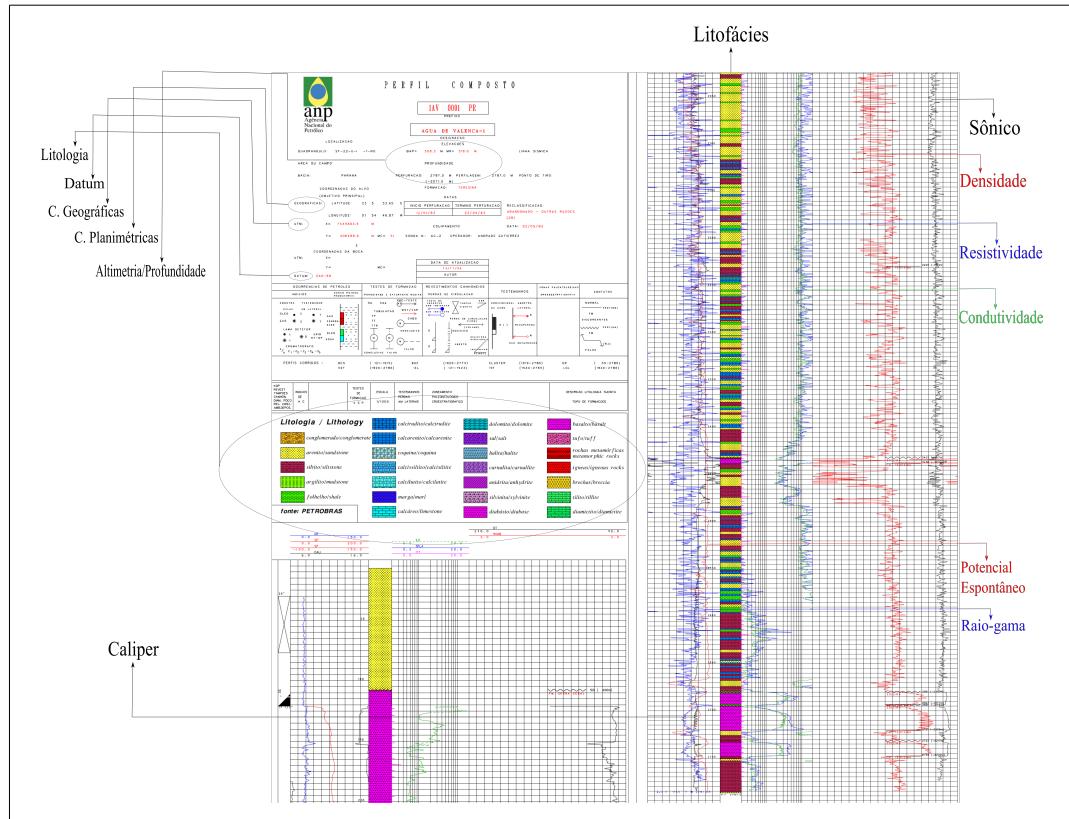


Figura 4.1: Exemplo de um dado público de uma perfilagem de poço composta realizada pela Petrobras, na Bacia do Paraná.

Entretanto não é toda a perfilagem de poço que contém o topo e base de camada. As RNAs se apresentam como uma solução para o problema de identificação litológica e dos topos e bases dessas camadas. Uma vez observado que a variação das propriedades físicas das rochas em subsuperfície variam obedecendo certos padrões (YAN *et al.*, 2014).

Em TELFORD e SHERIFF (1993), encontram-se variações das propriedades físicas dos principais grupos de rochas. A Tab. 4.1 e a Tab. 4.2 apresentam um compêndio desses principais valores.

Tabela 4.1: Compilação de Perfis usados na inferência de litologia.

Rocha	Densidade ( $g/cm^3$ )	Raios-Gama ( $Ci/g$ )	Potencial-Espontâneo ( $mV$ )
Conglomerado	2,50	—	—
Arenito	2,35	2,00 $\leftrightarrow$ 4,00	—
Folhelho	2,40	—	—
Argilito	2,55	—	—
Siltito	2,21	—	—
Dolomita	2,70	8,00	—
Marga	2,50	—	—
Basalto	2,99	0,50	—
Diabásio	2,90	—	—
Lava	2,61	0,33	—
Granito	2,64	0,70 $\leftrightarrow$ 4,80	—
Gabro	3,03	—	—
Peridotito	3,15	—	—
Quartzito	2,60	5,00	—
Xisto	2,64	—	—
Gnaisse	2,80	—	—
Serpentinito	2,78	—	—
Anfibolito	2,96	—	—
Eclogito	3,37	—	—
Mármore	2,75	—	—

Tabela 4.2: Compilação de Perfis usados na inferência de porosidade, permeabilidade.

Rocha	Resistividade ( $\Omega/m$ )	Neutrão (API)	Velocidade (km/s)
Conglomerado	$2 \times 10^3 \leftrightarrow 10^4$	—	$1,80 \leftrightarrow 4,90$
Arenito	$1 \leftrightarrow 6,4 \times 10^8$	—	$4,00 \leftrightarrow 4,30$
Folhelho	$50 \leftrightarrow 10^7$	—	$2,15 \leftrightarrow 3,30$
Argilito	$10 \leftrightarrow 8 \times 10^2$	—	—
Siltito	$1 \leftrightarrow 100$	—	$4,00 \leftrightarrow 6,20$
Dolomita	$3,5 \times 10^2 \leftrightarrow 5 \times 10^3$	—	$5,70 \leftrightarrow 6,00$
Marga	$3 \leftrightarrow 70$	—	—
Basalto	$10 \leftrightarrow 1,3 \times 10^7$	—	$5,00 \leftrightarrow 5.80$
Diabásio	$20 \leftrightarrow 5 \times 10^7$	—	—
Granito Porfirítico (seco)	$1,3 \times 10^6$	—	5,80
Granito Porfirítico (úmido)	$4,5 \times 10^3$	—	$5,00 \leftrightarrow 5.60$
Gabro	$10^3 \leftrightarrow 10^6$	—	$5,00 \leftrightarrow 5.80$
Peridotito (seco)	$6,5 \times 10^3$	—	—
Peridotito (úmido)	$3 \times 10^3$	—	—
Xisto	$20 \leftrightarrow 10^4$	—	—
Gnaisse (seco)	$3 \times 10^6$	—	—
Gnaisse (úmido)	$6,8 \times 10^4$	—	—
Tufa (seca)	$2 \times 10^3$	—	$1,80 \leftrightarrow 3,50$
Tufa (úmida)	$10^5$	—	—
Mármore	$10^2 \leftrightarrow 2,5 \times 10^8$	—	—

## 4.1 Modelo proposto para gerar os dados sintéticos

O modelo proposto para o teste da rede neuronal foi concebido com base em um modelo geológico esquemático proposto por MOHRIAK *et al.* (2008) *apud* (EIRAS, 1996). Esta simplificação reproduz, em uma bacia do tipo sinéclise, estruturas geológicas como Horts, Grábens, semi-grábens, falhas normais, reversas. E representa ainda processos halocinéticos referenciados por EIRAS (1996).

A Fig. 4.2 representa o modelo em tratado anteriormente. A caixa aumentativa evidencia a falha normal, representada no dado de poço por um contato não plano-paralelo.

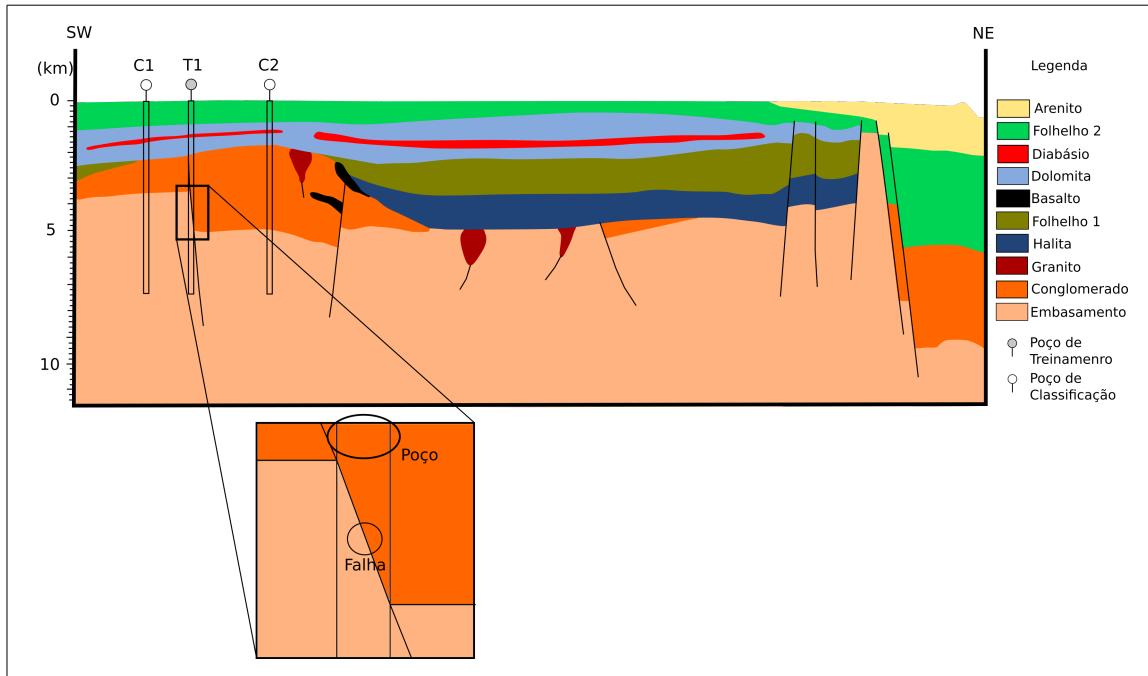


Figura 4.2: Modelo Simplificado baseado em MOHRIAK *et al.* (2008).

A partir da Fig. 4.2 foram gerados três poços, na parte SW do perfil, com profundidades de 7 km cada. Os três poços contém um conjunto com 4 dados de propriedades físicas que são densidade, raio-gama, resistividade e velocidade, respectivamente. Os valores de propriedades físicas utilizados foram baseados, em resultados já publicados, na literatura geocientífica, anteriormente, e retirados de TELFORD e SHERIFF (1993). Os poços simulam dados de *well logging* com uma taxa de amostragem de 10 m.

Os poços simulam diferentes padrões interpretativos usuais da ciência de perfilação<sup>1</sup>. O poço denominado T1<sup>2</sup> se localiza entre os poços C1 e C2 atravessando uma falha normal.

A Fig. 4.3 apresenta os dados do poço T1. As espessuras das camadas são de 800 m de embasamento, 2 km de uma mistura crescente entre conglomerado e embasamento, perfazendo um padrão sino nos dados de perfilagem, 2 km de conglomerado, 1 km de dolomita (pacote inferior), 300 m de diabásio, 400 m de dolomita (pacote superior), 600 m de folhelho 2. A falha foi representada por uma função linear da variação de profundidade por propriedade física de uma mistura crescente de conglomerado e embasamento.

<sup>1</sup>Padrões usuais reconhecido por intérpretes geralmente associados a horizontes de interesse. Esses padrões são identificados como padrões sinos, sinos invertidos, serra e caixa.

<sup>2</sup>T1: poço escolhido para treinar a rede neuronal.

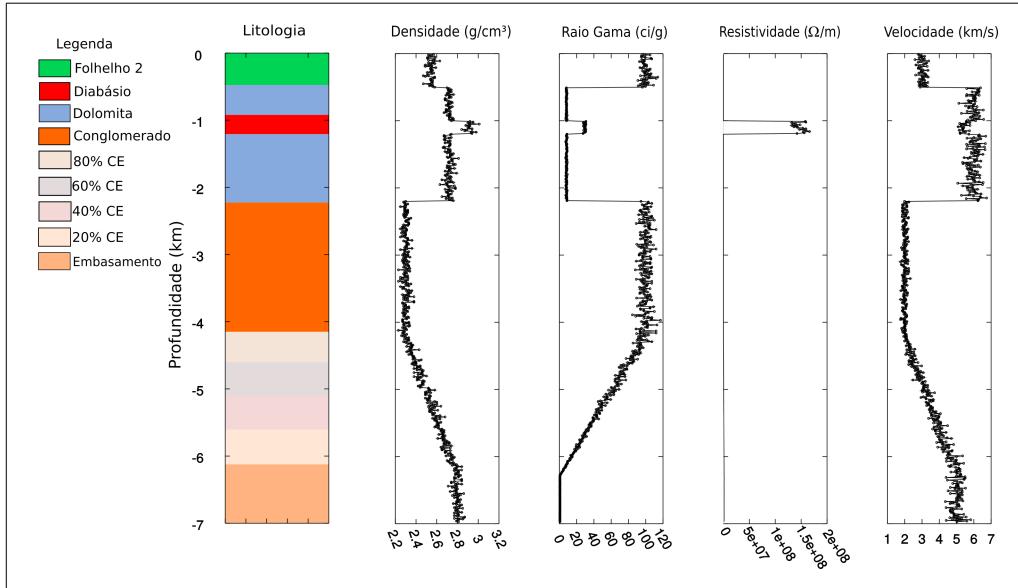


Figura 4.3: Dado de perfilagem sintético, T1. Aonde a porcentagem de CE indica a mistura de conglomerado com embasamento

O poço C1<sup>3</sup>, Fig. 4.4, possui as mesmas classes de rochas do poço T1. A escolha da posição dos poços quase que exclusivamente na parte SW do perfil se deu em virtude da localização do poço T1. Uma vez que espera-se da rede já treinada um reconhecimento das classes já estudadas. Os pacotes sedimentares apresentam espessuras de 2,8 km de embasamento, 1,6 km de conglomerado, 1 km de dolomita (segundo pacote), 200 m de diabásio, 500 m de dolomita (primeiro pacote) e 500 m de folhelho.

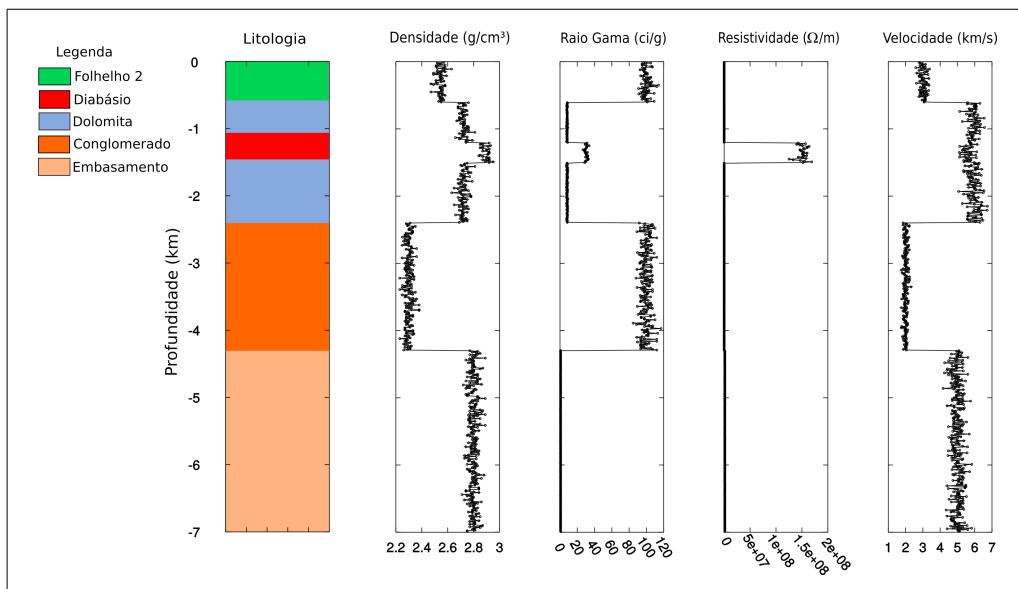


Figura 4.4: Dado de perfilagem sintético, C1.

<sup>3</sup>C1: Poço de classificação da rede neuronal número 1.

O poço C2<sup>4</sup>, Fig. 4.5, localiza-se em um alto estrutural, e apresenta espessura de 5 km de conglomerado. O embasamento possui uma espessura de 1,8 km. Os pacotes de folhelho 2, dolomita (pacote superior) e diabásio 500 m respectivamente. E o segundo pacote sedimentar de dolomita 1,6 km.

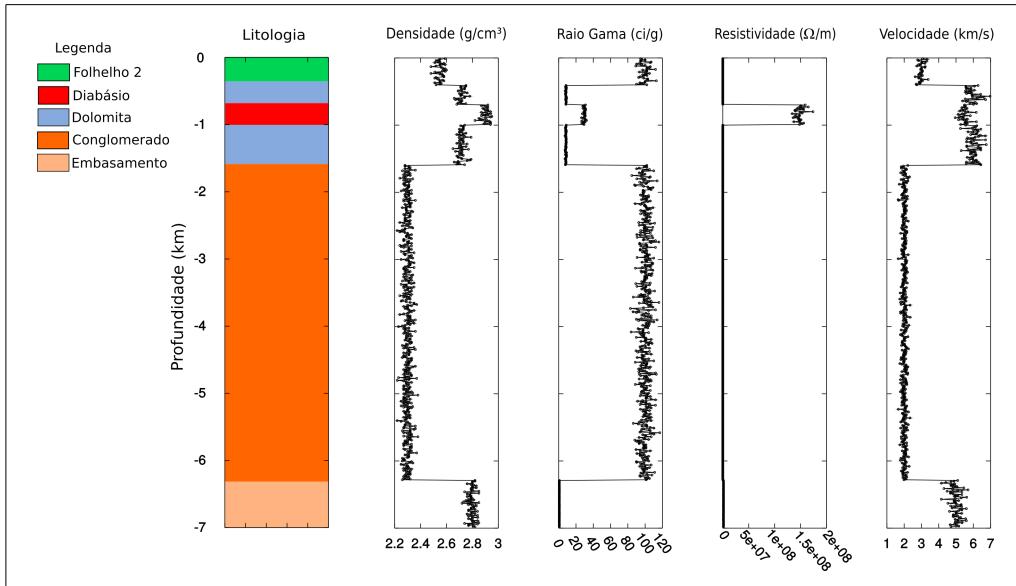


Figura 4.5: Dado de perfilagem sintético, C2.

## 4.2 Dado Real

A triagem dos dados públicos contemplaram 506 arquivos \*.dlis, 113 \*.lis, 118 dados adicionais, 125 perfis compostos digitalizados, 174 poços públicos, 120 arquivos \*.agp, todos localizados na Bacia Sedimentar do Paraná.

O conjunto de dados \*.lis e \*.dlis estão sendo convertidos para arquivos em formato texto, que serão posteriormente concatenados com os arquivos \*.agp afim de se obter o input da rede. Este processo ainda se encontra na fase inicial com cerca de 3% concluído.

A Fig. 4.6 mostra a localização e distribuição dos poços na Bacia do Paraná.

---

<sup>4</sup>C2: Poço de classificação da rede neuronal número 2.

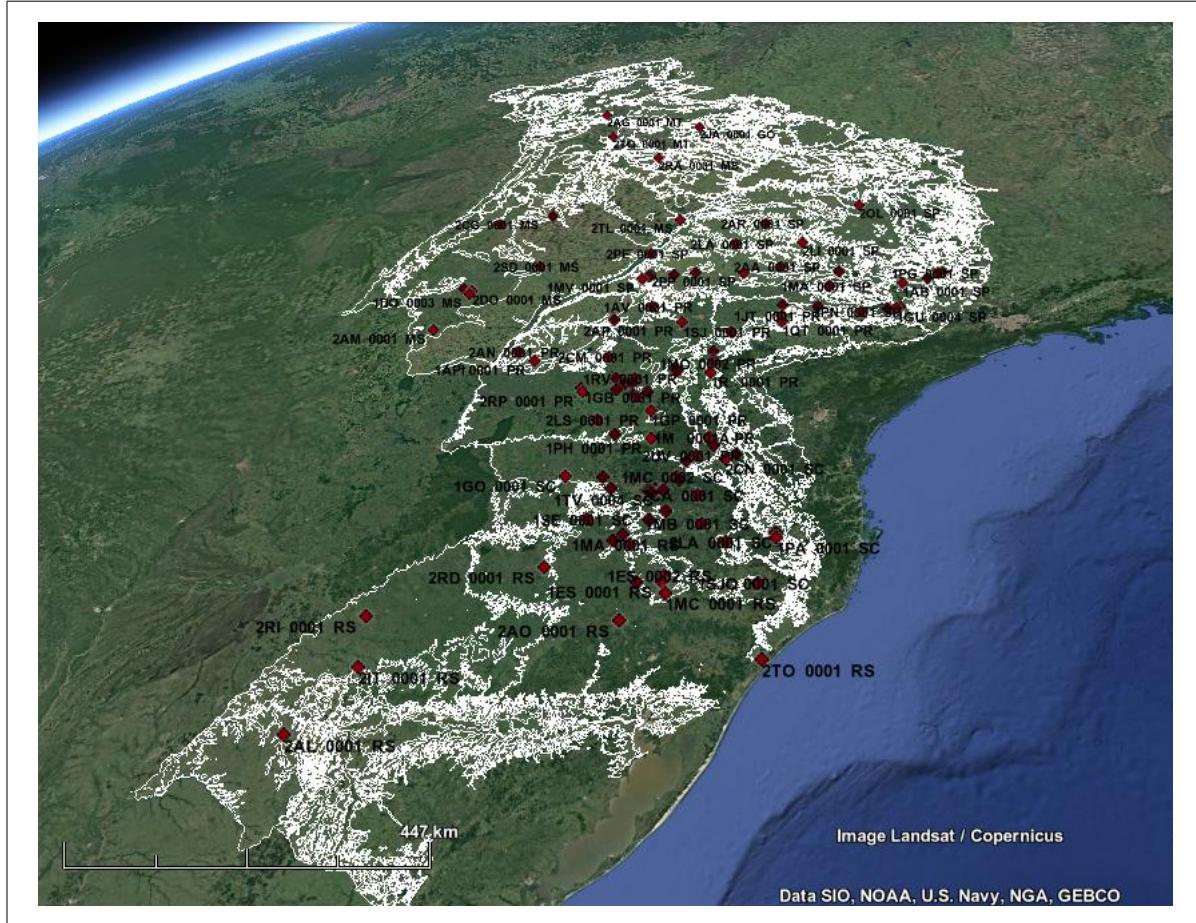


Figura 4.6: Localização do total de poços de trabalho.

#### 4.2.1 Dados reais e treinamento e classificação da rede

Como uma primeira abordagem com os dados reais foram escolhidos dois poços localizados na borda leste da Bacia Sedimentar do Paraná que prenchessem alguns pré-requisitos.

1. Os primeiros poços devem estar localizados próximos para garantir um controle inicial no que tange as litologias presentes no poço de treinamento.
2. O poço de treinamento deve possuir a maior quantidade possível de dados por litologia
3. As entradas da rede devem ser compostas das mesmas propriedades físicas
4. As propriedades físicas que definem litologia devem ser priorizadas quando possível.

O mapa da Fig. 4.7 aponta a localização dos dois poços iniciais para o treinamento da rede.

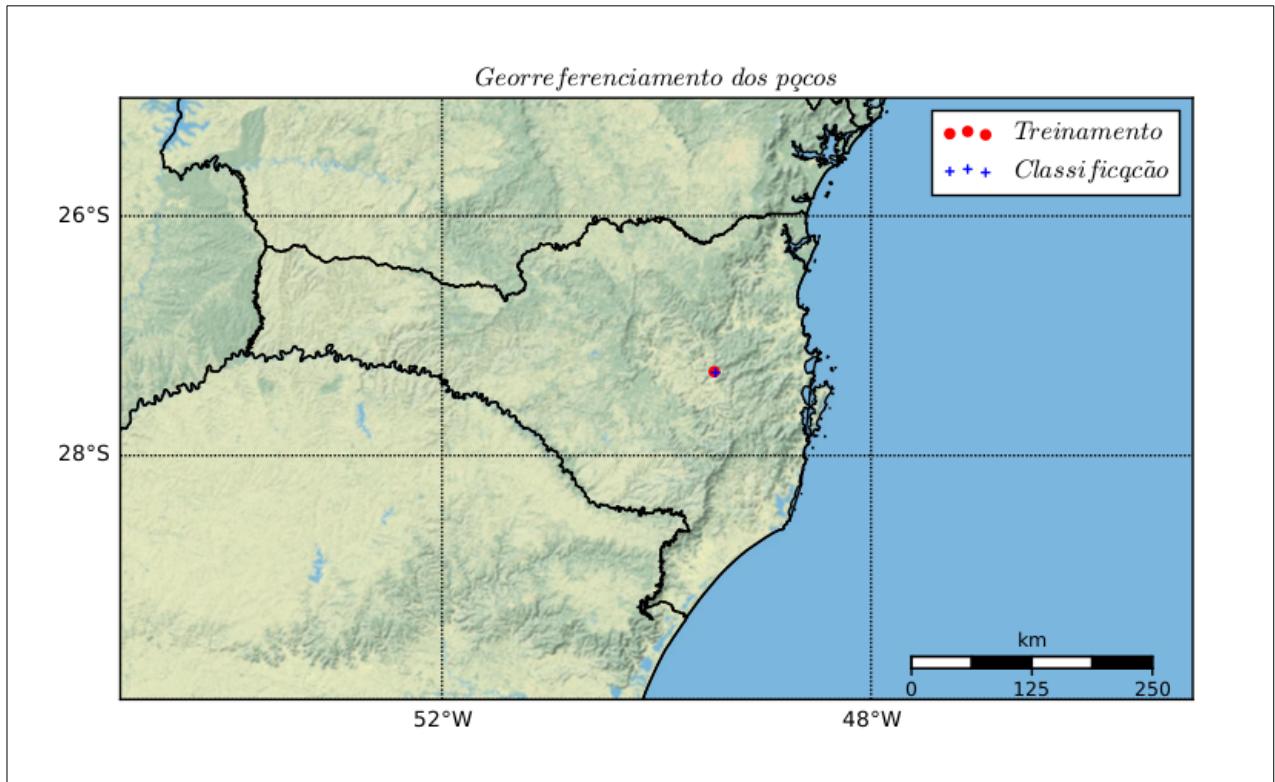


Figura 4.7: Localização dos poços escolhidos.

O primeiro poço escolhido para o treinamento da rede foi o de nome 1BN0001SC localizado na borda leste da Bacia Sedimentar do Paraná no Estado de Santa Catarina. A taxa de amostragem é 0,2 medidas por metro e possui uma profundidade de aproximadamente 1400 m. As propriedades medidas foram o potencial espontâneo, raios-gama, resistividade lateral, radiação neutrônica além de medidas de propriedades para inferência da qualidade do poço e medidas de complementação a perfis sísmicos tais como *Transient Time Integrator*.

O segundo poço denomina-se 1BN0002SC e foi perfurado a poucos quilômetros de distância ao sul do poço de treinamento. Possui uma profundidade de 700 m e taxa de amostragem no valor de 0,2 medidas por metro. As propriedades inferidas foram Potencial espontâneo, resistividade lateral, densidade de formação, raio-gama, além de medidas de propriedades para inferência da qualidade do poço e medidas de complementação a perfis sísmicos tais como *Integral Time Travel*.

Foram inicialmente escolhidas quatro propriedades que são comuns em ambos os poços dentre as quais podem-se destacar o potencial espontâneo, resistividade lateral, TTI, TOT. As Figuras 4.8 e 4.9 apresentam a distribuição de propriedades físicas relacionadas com a litologia associado à variação de profundidade, respectivamente.

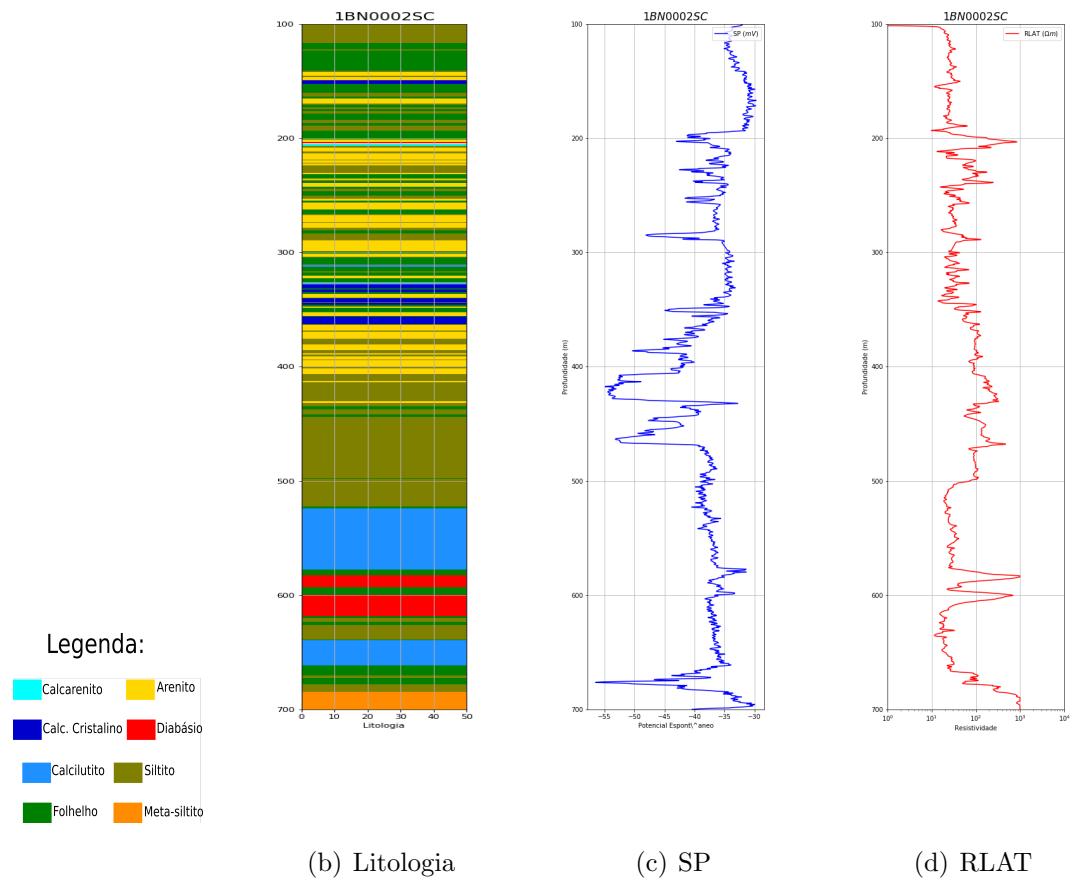


Figura 4.8: Poço 1BN0002SC e as respectivas propriedades físicas escolhidas para o teste. Em (b) tem-se a variação de litologia com a profundidade, (c) o potencial espontâneo e em (d) a resistividade lateral.

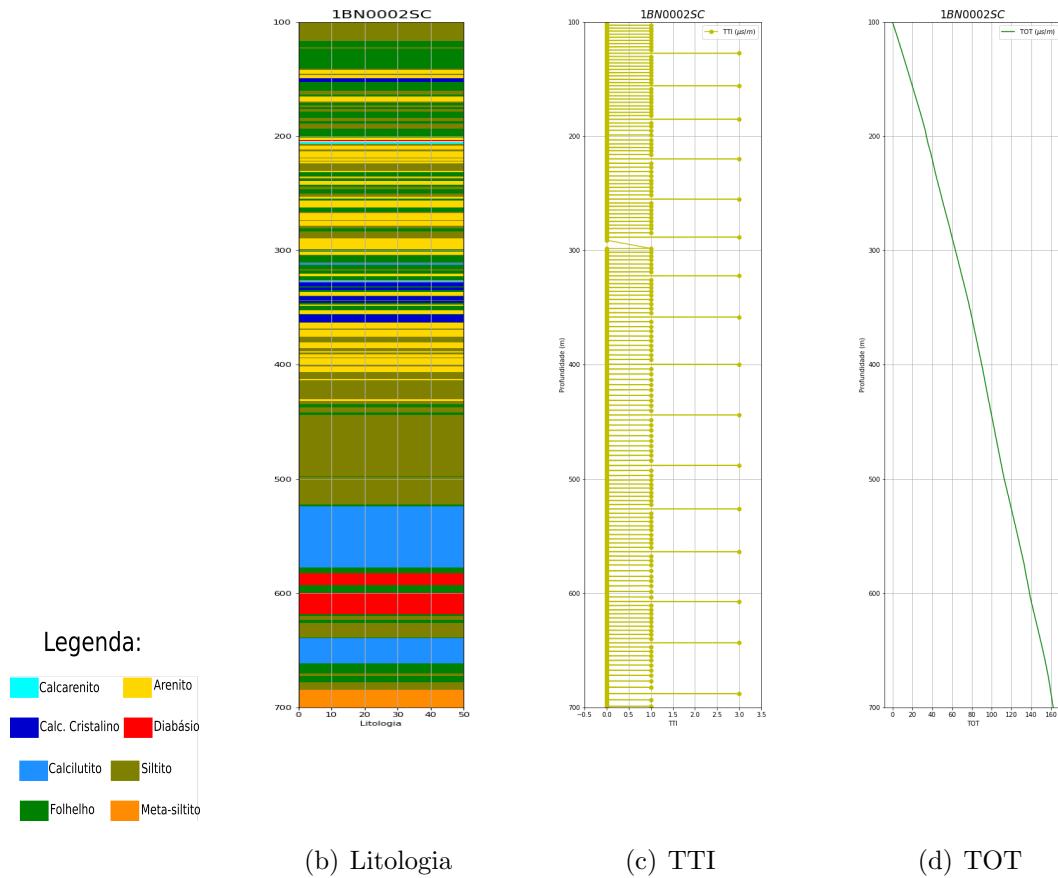


Figura 4.9: Poço 1BN0002SC e as respectivas propriedades físicas escolhidas para o teste. Em (b) tem-se a variação de litologia com a profundidade, (c) *Transient Time Integrator* e em (d) TOT.

Muito embora o critério de escolha dos poços tenham sido obedecidos alunas litologias presentes no poço de classificação não existem no poço de treinamento. Isso torna necessário que futuramente seja criado um banco de dados para classificação de rochas que contenham dados de propriedades que sejam independentes da profundidade. Extinguindo-se o conceito poço de treinamento substituindo-o por Banco de dados litológicos.

# Capítulo 5

## Resultados e Discussões

A Fig. 5.1 apresenta à análise de agrupamentos das propriedades físicas analisadas por classes de rochas para o poço T1. Foi utilizado para tal foi adotado um padrão de cores para cada tipo litológico.

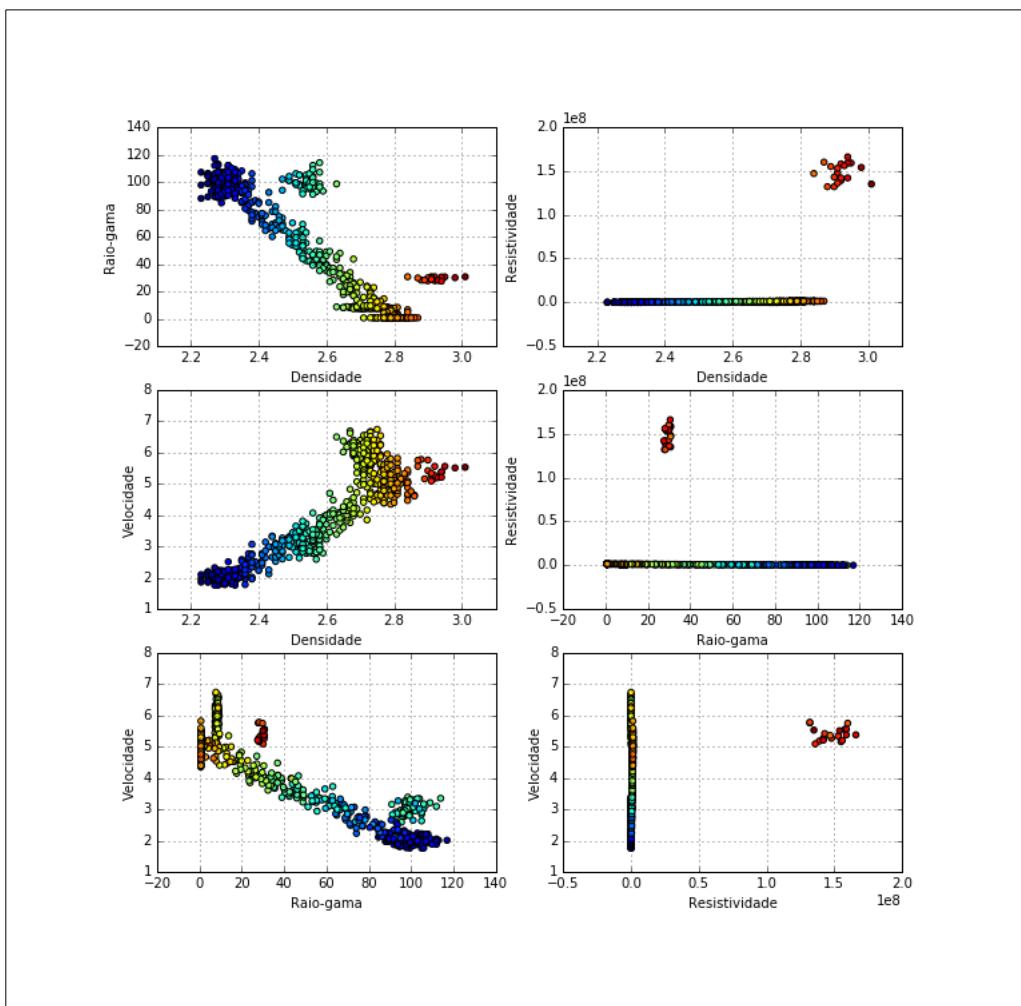


Figura 5.1: Agrupamento de dados do poço T1.

Em vermelho escuro, se encontra o diabásio, a graduação de cores entre o vermelho

claro e o amarelo, se encontra o embasamento, a gradação de cores entre o laranja e o verde claro encontra-se a dolomita, verde claro se encontra o folhelho 2, a gradação de azul para azul escuro encontra-se o conglomerado, e a gradação que vai do amarelo ao azul são as subclasses de mistura de conglomerado com embasamento de 20%, 40%, 60% e 80%, respectivamente.

É perceptível o notável contraste de variação de resistividade entre a rocha de origem ígnea, em contraste com as propriedades físicas das demais rochas de origem sedimentar e metamórfica. O agrupamento das rochas sedimentares formam um conjunto quase linear próximo a zero.

A Fig. 5.2 apresenta à variação das propriedades físicas analisadas por agrupamento de classes de rochas para o poço C1.

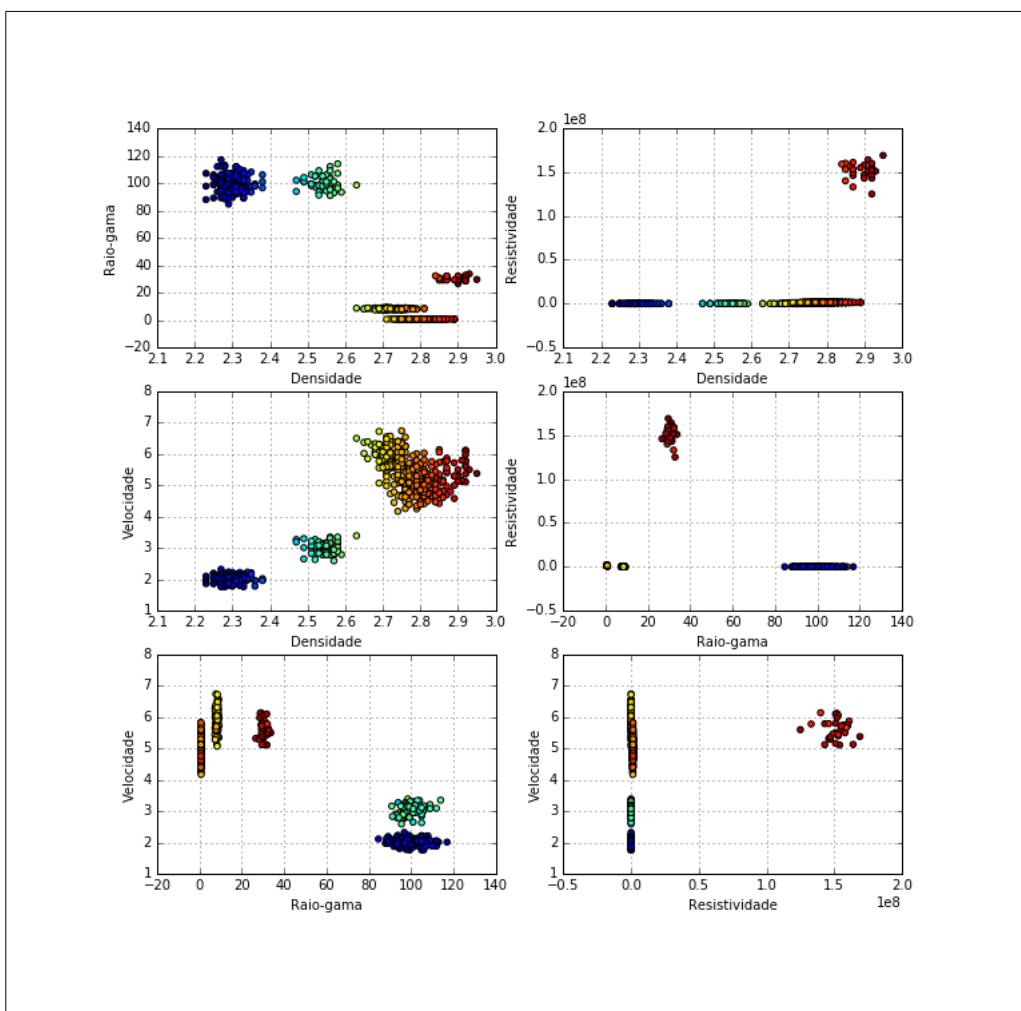


Figura 5.2: Agrupamento de dados do poço C1.

Neste caso, o agrupamento das classes de rochas é mais evidente, no gráfico de raio-gama por densidade, que evidencia os 5 litotipos distintamente. E, da mesma maneira, o gráfico de velocidade por densidade.

A Fig. 5.3 apresenta à variação das propriedades físicas analisadas por agrupa-

mento de classes de rochas para o poço C2. Em destaque, de vermelho, o litotipo diabásio.

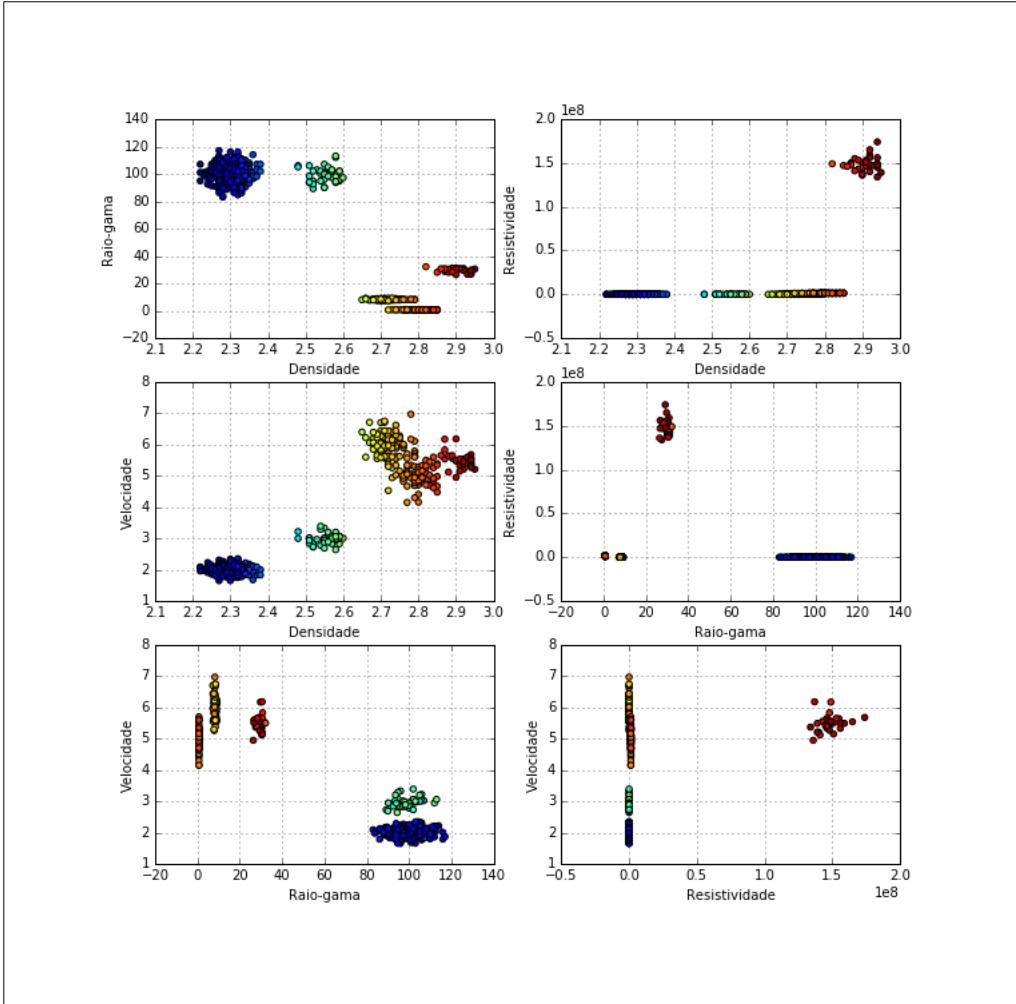


Figura 5.3: Agrupamento de dados do poço C2.

Na mesma forma, o agrupamento das classes de rochas é mais evidente, no gráfico de raio-gama por densidade, que evidencia os 5 litotipos distintamente. E, da mesma maneira, o gráfico de velocidade por densidade.

## 5.1 Treinamento

A etapa de treinamento consiste em um ajuste de pesos dos neurônios da rede. Nesta fase, é identificado o neurônio que tem os valores dos pesos mais parecidos com os parâmetros de entrada da rede. Por conseguinte, os diversos mapas são obtidos através dos sucessivos ciclos de treinamento ao longo do tempo. A Fig. 5.4(a), representa a organização da rede com apenas um ciclo de treinamento. Nesta imagem, a rede ainda não é capaz de identificar nenhuma litologia. Ao se aumentar o número de ciclos é perceptível que o ajuste dos pesos cria um conjunto de neurônios

vencedores capazes de identificar as classes litológicas. Na quinta iteração, Fig. 5.4(b), as classes folhelho 2 e dolomita, por exemplo (cores mais azuis) ocupam a maior área do mapa. Já na milésima iteração, Fig. 5.4(d), a área azul é reduzida dando lugar as cores amarela e verde, que representam as subclasses de conglomerado e embasamento.

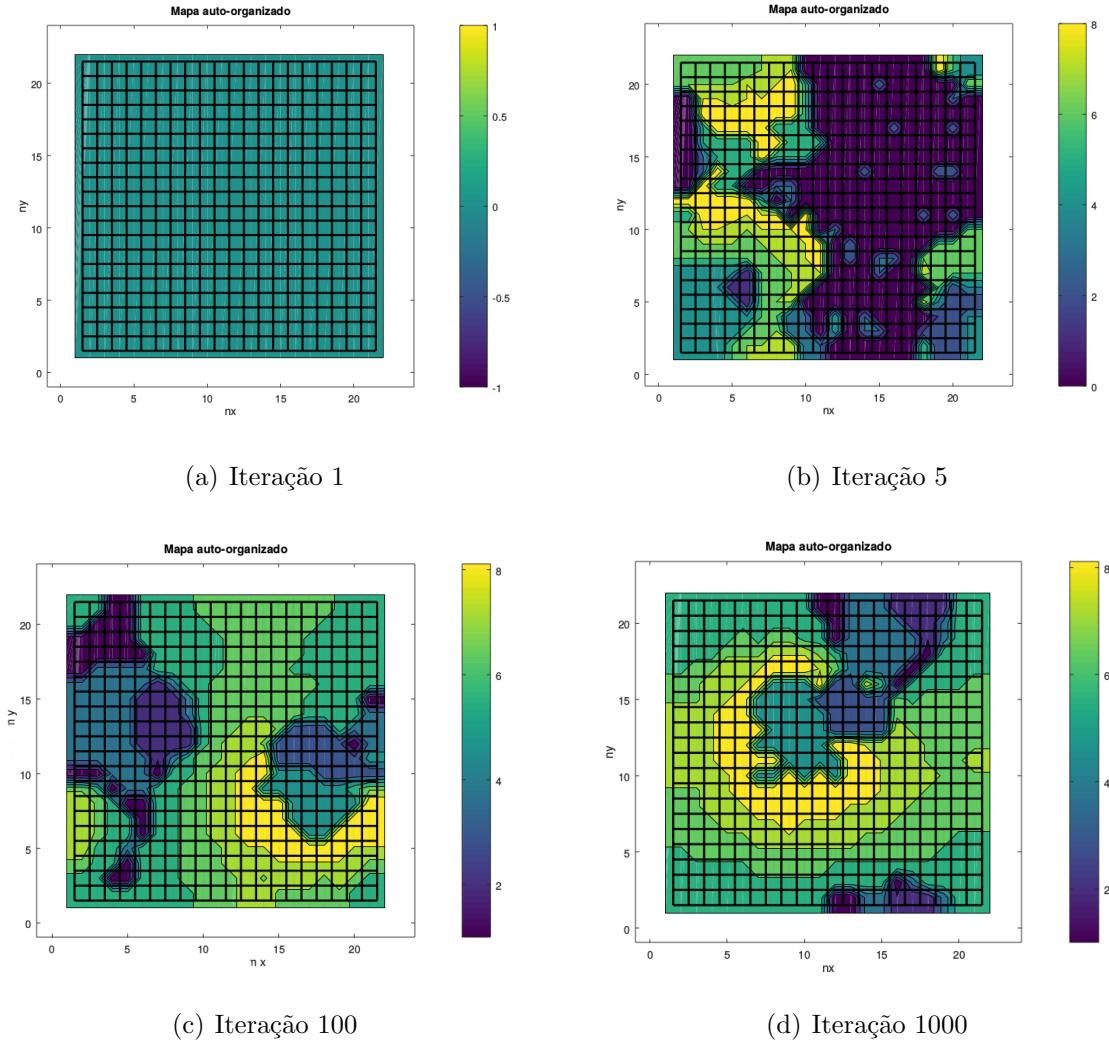


Figura 5.4: Mapas auto-organizáveis e sua evolução temporal.

Os mapas da Fig. 5.4 apresentam as zonas do hiperplano especializadas em identificar as classes de rochas. O código numérico 1 representa folhelho, 2 dolomita, 3 diabásio, 4 conglomerado, 5 embasamento, 6 mistura conglomerado/embasamento 80%, 7 mistura conglomerado/embasamento 60%, 8 mistura conglomerado/embasamento 40% e 9 mistura conglomerado/embasamento 20%. A Tab. 5.1 faz um paralelo entre o código numérico utilizado com *output* da rede e as litologias do modelo.

Litologia	Código numérico
Folhelho 2	1
Dolomita	2
Diabásio	3
Conglomerado	4
Conglomerado 80%	5
Conglomerado 60%	6
Conglomerado 40%	7
Conglomerado 20%	8
Embasamento	9

Tabela 5.1: Tabela de referência para conversão do padrão numérico em litologia.

A Fig. 5.5 apresenta o teste de convergência da rede neuronal.

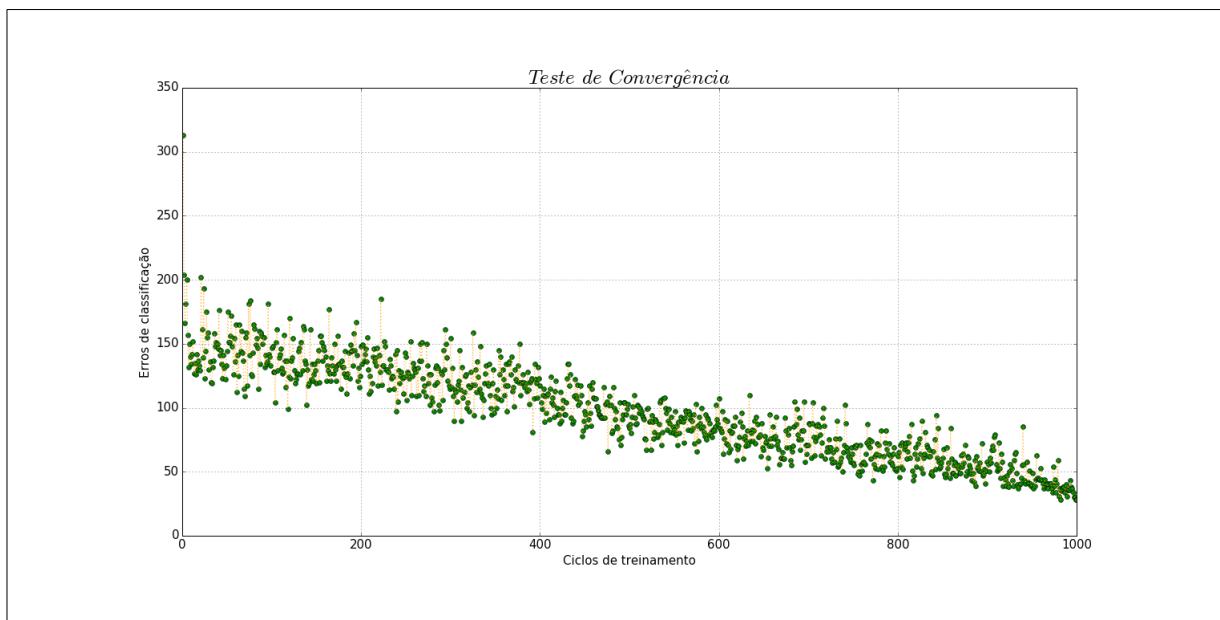


Figura 5.5: Teste de convergência da rede.

O teste de convergência é realizado durante a fase de treinamento e mostra que a rede se encontra estabilizada em 1000 ciclos de treinamento com 28 erros de classificação, ou seja, um erro de 4%. Isto significa ser inócuo aumentar a iteração afim de diminuir o erro.

## 5.2 Identificação

A seguir são apresentados os resultados da etapa de classificação da rede foram acrescentados os resultados dos classificadores euclidianos e de Mahalanobis. Nesta

fase, dois poços foram utilizados chamados de poços C1 e C2. O primeiro destes localizado a SW do perfil, Fig. 4.2, possui 7km de profundidade. A saída da rede, para o poço C1 está localizada ao lado direito da Fig. 5.6. Ao lado esquerdo é apresentada o poço original. Abaixo é mostrado uma breve estatística deste processo de identificação da rede. Ao lado esquerdo do poço identificado pela rede estão os resultados obtidos pelos classificadores euclideano e de Mahalanobis.

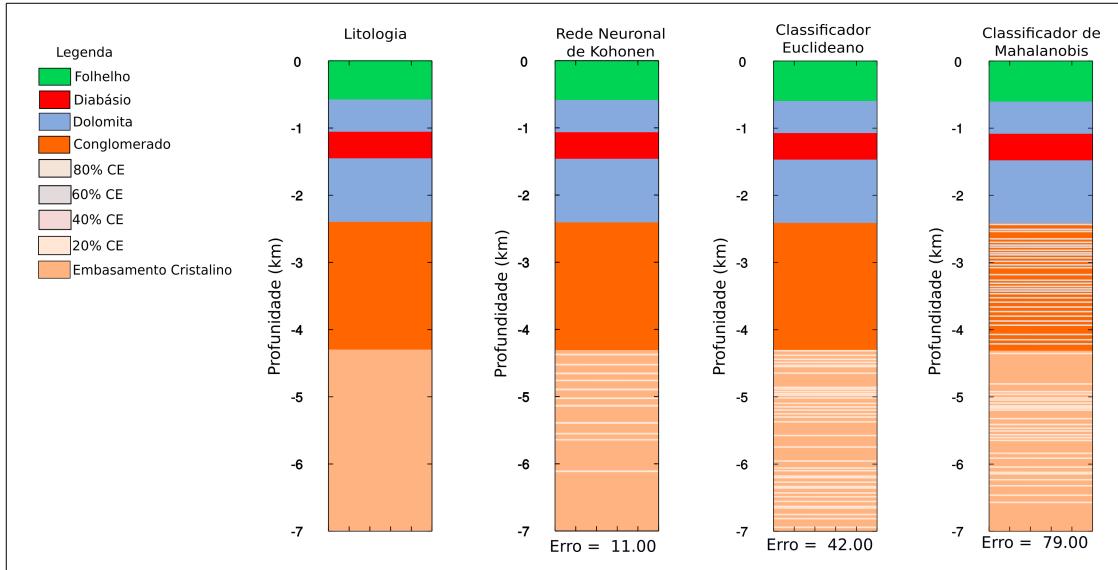


Figura 5.6: Dado de saída da rede para o poço de classificação C1.

O processo de identificação foi repetido, contudo, desta vez, para o caso do poço C2. Este localiza-se mais a NE do perfil, Fig. 4.2, no topo de um alto estrutural com igual profundidade de 7 km. A saída da rede, para o poço C2 está localizada ao lado direito da Fig. 5.7. Ao lado esquerdo é apresentada o poço original. Abaixo é mostrado uma breve estatística do processo de identificação. Ao lado esquerdo do poço identificado pela rede estão os resultados obtidos pelos classificadores euclideano e de Mahalanobis.

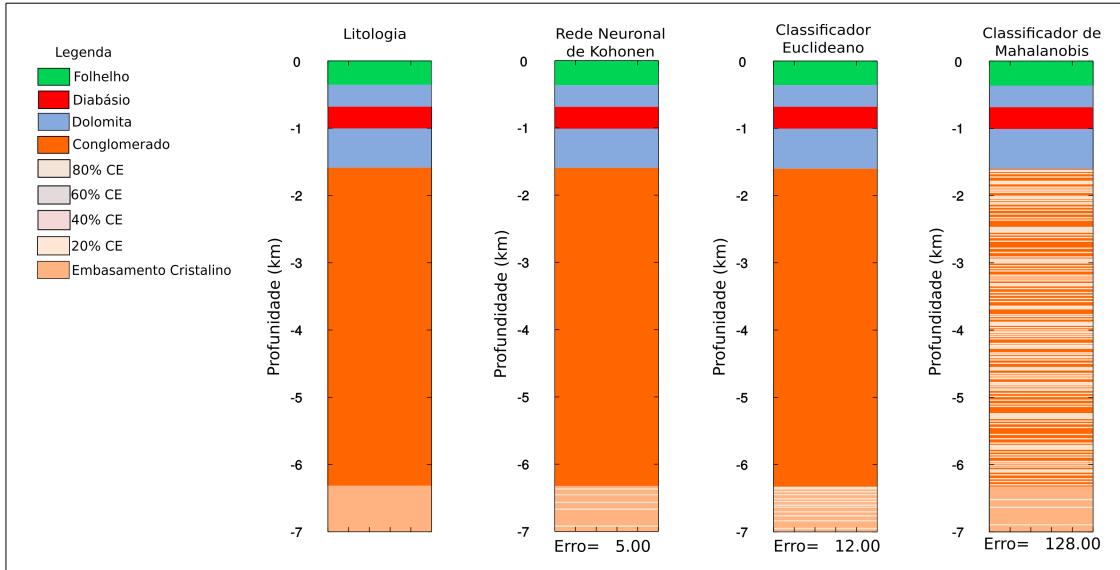


Figura 5.7: Dado de saída da rede para o poço de classificação C2.

Em ambos os casos de identificação, o número de neurônios vitoriosos igualou-se ao total de neurônios da rede, Tab. 5.2. Isto indica o máximo de aproveitamento durante os processos, com um tempo de máquina atingindo 25 segundos.

Tabela 5.2:

## Estatística da Rede

Dados	Poço C1	Poço C2
Dados de treinamento	697	697
Dados a serem classificados	699	698
Neurônios da Rede	400	400
Neurônios vitoriosos	400	400
Neurônios sem uso	0	0
Erro	11	5

Os classificadores apresentaram dois comportamentos distintos. No poço C1, o classificador de Euclides apresentou 42 erros confundindo rochas do embasamento com rochas com 20% de conglomerado e embasamento. No poço C2, o classificador de Euclides apresentou um maior número de erros 12 associados ao embasamento cristalino classificando alguns pontos como conglomerado com 20% de conglomerado e embasamento.

Em contra-partida, o classificador de Mahalanobis apresentou 79 erros, no total do poço C1 trocando as rochas do embasamento e conglomerado por dois tipos específicos de rocha: as rochas do 20% e 60% CE. E apresentou 128 erros

bem distribuídos ao longo das rochas do embasamento e conglomerado, no poço C2 confundindo-as com rochas de 60% e 20% de conglomerado com embasamento.

### 5.3 Dado Real: treinamento

Foram performados ao todo 6 testes no que tange o dado real. Os parâmetros testados da rede neuronal foram o número de neurônios e épocas<sup>1</sup>. Analogamente ao dado sintético foram analisadas os mapas auto-organizáveis e as curvas de convergência da rede. Para cada teste foram gerados três mapas SOMs que representam o início o meio e o fim do processo de treinamento e aprendizagem da rede. O por fim as curvas de convergência apontam todo o processo de aprendizado da rede.

Como a taxa de amostragem do dado real é alta optou-se por aumentar a dimensão do hiperplano da rede. O teste 01 foi conduzido em uma malha composta por 1600 neurônios, 10 épocas, 976 neurônios vitoriosos e um custo computacional de 6,388s em tempo de máquina. o conjunto de mapas apresentados na Fig. 5.8

---

<sup>1</sup>Época é definido pelo número de iterações ao longo do tempo.

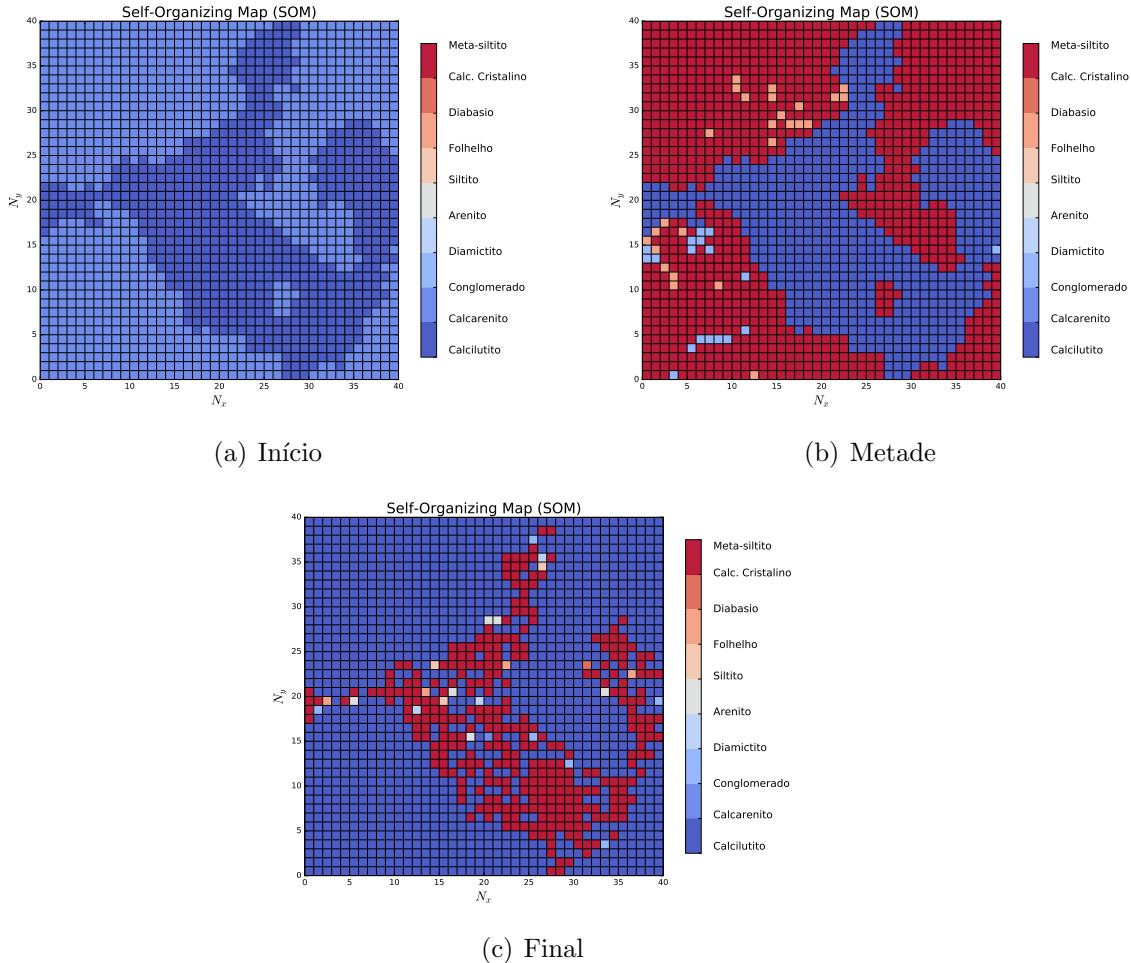


Figura 5.8: Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 40X40 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento.

O teste 02 foi conduzido em uma malha composta por 1600 neurônios, 100 épocas, 1600 neurônios vitoriosos e um custo computacional de 62,01s em tempo de máquina. o conjunto de mapas apresentados na Fig. 5.9

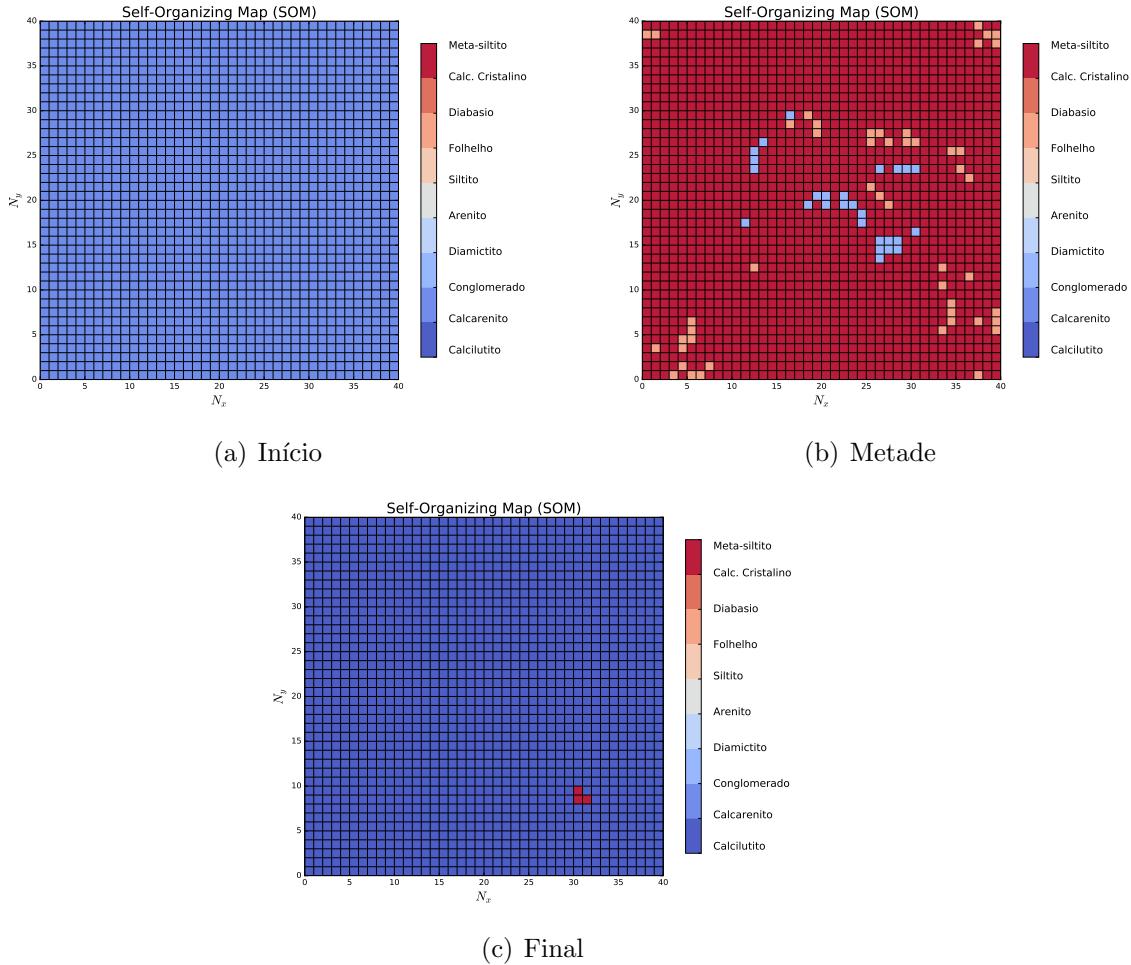


Figura 5.9: Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 40X40 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento.

O teste 03 foi conduzido em uma malha composta por 1600 neurônios, 1000 épocas, 1600 neurônios vitoriosos e um custo computacional de 610,12s em tempo de máquina. o conjunto de mapas apresentados na Fig. 5.10

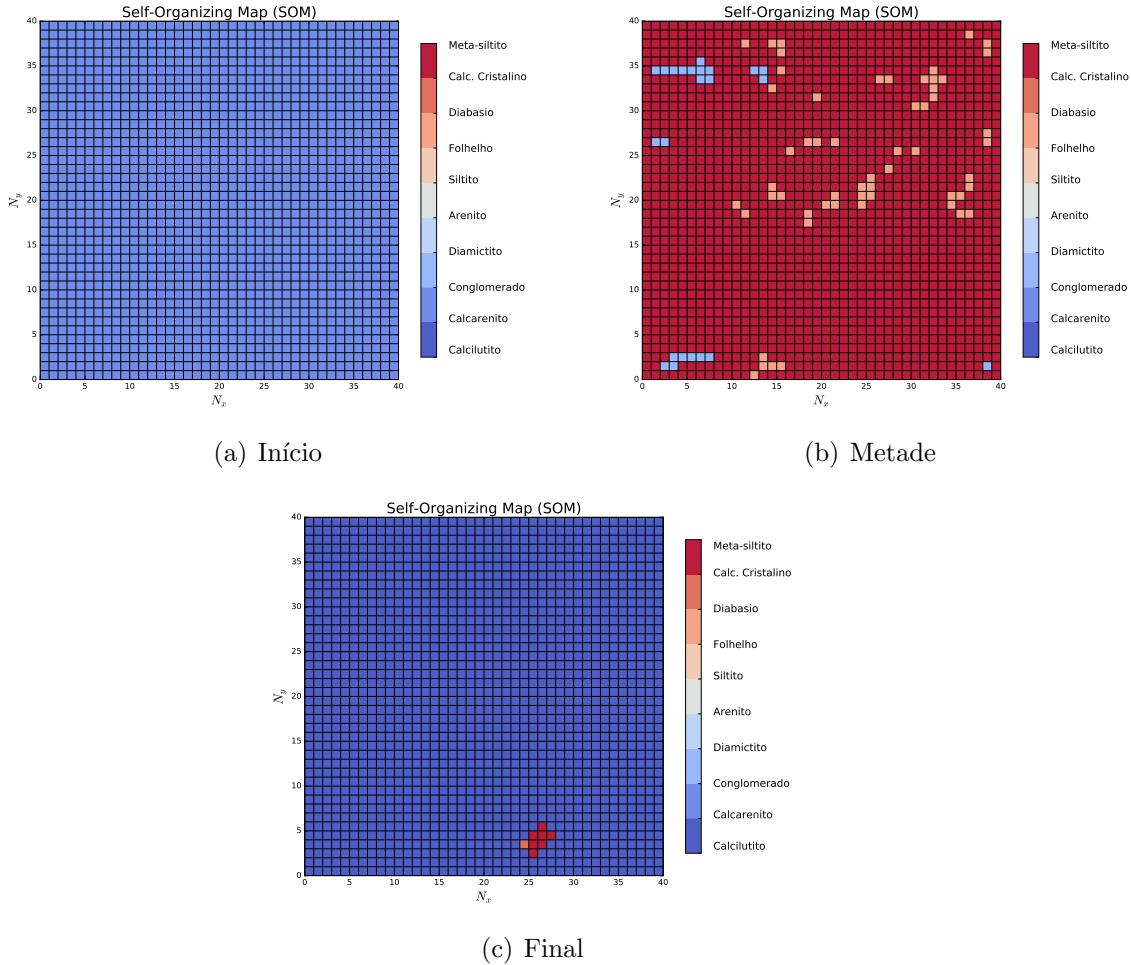


Figura 5.10: Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 40X40 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento.

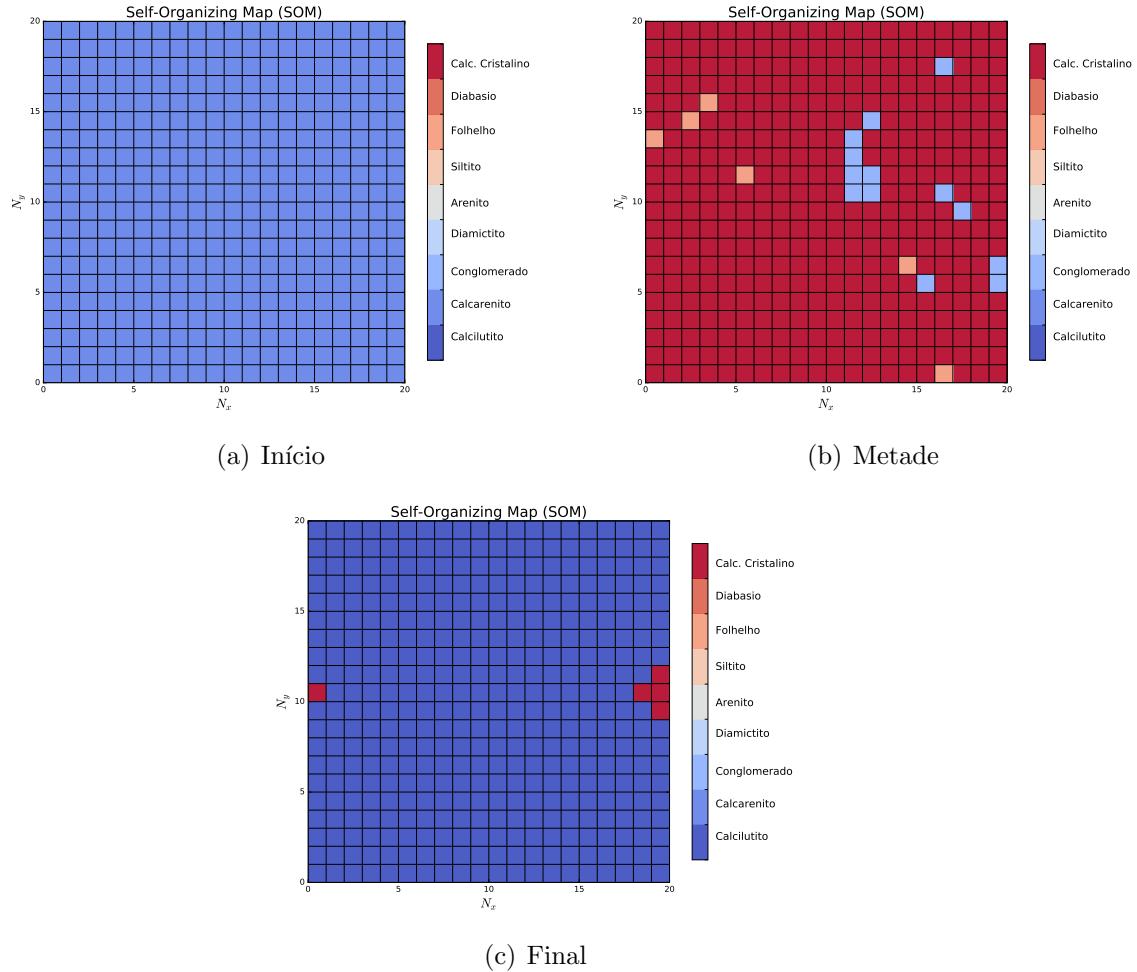


Figura 5.11: Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 20X20 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento.

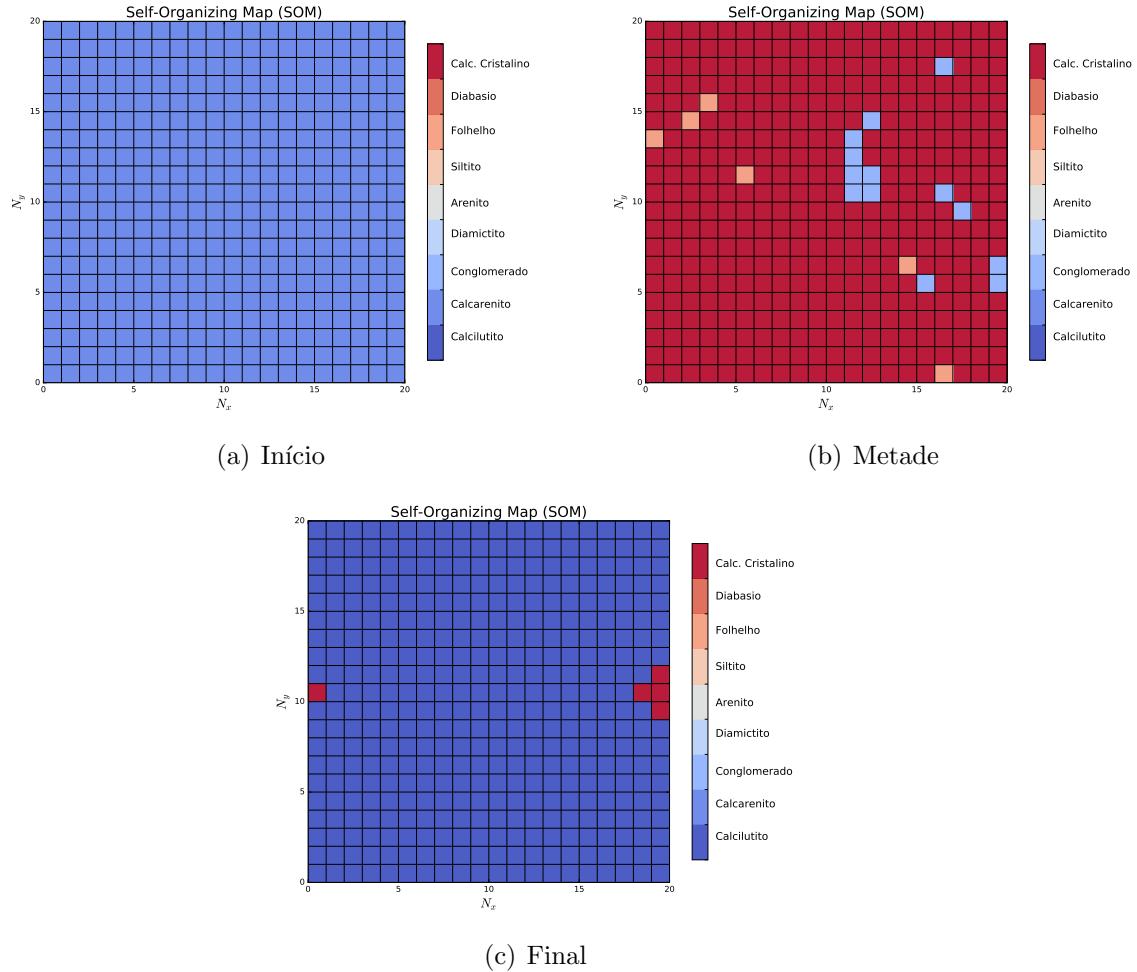


Figura 5.12: Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 20X20 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento.

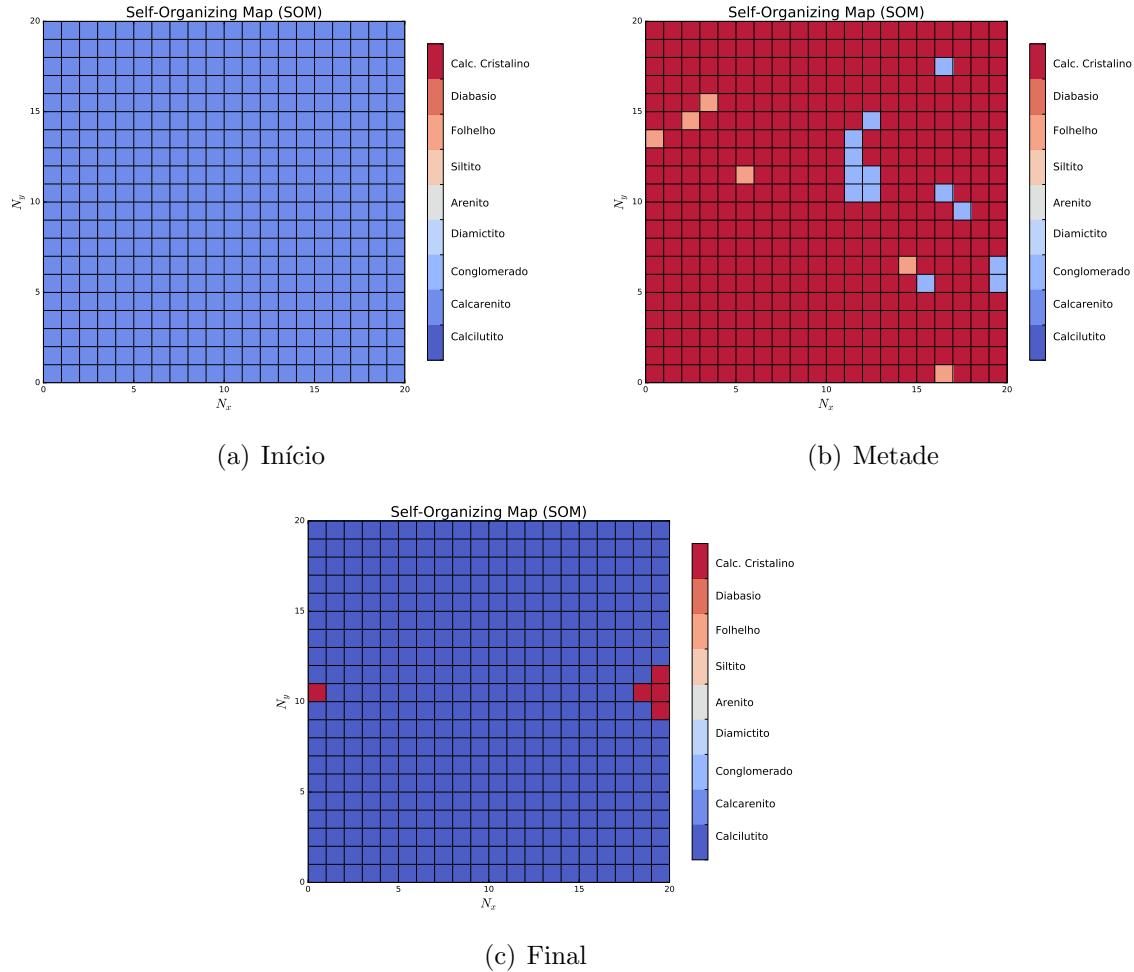


Figura 5.13: Mapas auto-organizáveis e sua evolução temporal. A figura (a) mostra a rede com 20X20 neurônios no início do processo de treinamento. A figura (b) apresenta a rede no meio do processo de treinamento e (c) a rede no final do processo de treinamento.

# Capítulo 6

## Conclusões

O teste de convergência da rede, Fig 5.5, realizado durante a etapa de treinamento, indicou que o número de erros não iria diminuir após o milésimo ciclo de treinamento. Sendo o resultado, Fig. 5.4, deste teste usado como parâmetro para o número de repetições realizadas para os casos de identificação da rede.

Os diagramas de velocidades por densidade e o de velocidade por raio-gama, Fig. 5.1, Fig. 5.2 e Fig. 5.3, apresentaram os agrupamentos mais bem separados. Portanto estas propriedades físicas (densidade, velocidade e raio-gama) tem uma importância relativa maior ,na classificação das litologias dos poços C1 e C2.

A saída da rede aponta que o maior caso de erros ocorreram em uma única classe de rocha, a do embasamento. Esses erros fizeram com que conglomerados fossem classificados como rochas do embasamento, nos dois casos dos poços de classificação, o poço C1 e o poço C2. Uma das razões pode ser o fato das misturas de conglomerado e embasamento serem finas demais para a rede conseguir realizar uma identificação de padrão. Ou pelo fato dos conjuntos de propriedades físicas da mistura de 20% se aproximar das propriedades físicas que representam o litotipo embasamento.

O menor número de erros relativos encontrados, no poço C2, Fig. 5.7, deve-se a escolha da alocação do furo, no perfil. O poço C2 localiza-se em um baixo estrutural, atingindo menos de 1km do embasamento. Entretanto, o poço C1, Fig. 5.6, encontra-se em um alto estrutural, divergindo do poço C2 e produzindo, consequentemente, os maiores erros relativos encontrados.

O classificador de Euclides apresentou mais erros do que a rede neuronal de Kohonen com 42 erros para o poço C1 e 12 erros para o poço C2. E o classificador de Mahalanobis apresentou o resultado de identificação de poços com os maiores erros 79 e 128 respectivamente para os poços C1 e C2.

Tal desempenho dos classificadores se deu por conta da existência de uma falha normal aonde foi escolhida a alocação do furo, Fig. 4.2. Nesta situação simulada há uma mistura entre os clusters em todos os espaços bi-dimensionais de propriedades analisadas tanto para o conglomerado quanto para o embasamento cristalino, Fig.

5.1. Portanto a definição dos centroides dos agrupamentos de propriedades ficam longe das distribuições ideais preditas no teste analítico do capítulo 1, na seção 1.4.3.

# Capítulo 7

## Cronograma

Ao longo do ano de 2017 cursei as disciplinas de Métodos Numéricos de On-das Sísmicas, no Observatório Nacional, no segundo trimestre. Durante o terceiro trimestre, cursei disciplinas externas Redes Neuronais, do Instituto de matemática da Universidade do Estado do Rio de Janeiro, ministrada pela Professora Roseli Widemann, que foi cursada semestralmente. E na Universidade Federal do Rio de Janeiro iniciei a disciplina de Aprendizado de Máquina do setor de Engenharia Elétrica Código CPE-775, mas tranquei ao julgar que ela não traria nenhuma contribuição ao meu projeto de doutorado.

Neste ano eu já me encontro inscrito na disciplina de Redes Neurais I código ELE2394 da PUC-Rio e ainda pretendo cursar a disciplina de sismologia que será oferecida ao longo desse ano no Observatório Nacional, bem como a disciplina de Inversão Não-linear.

Em **vermelho** encontra-se o mês atual.

Etapa	Meses																									
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24		
Pesquisa na Literatura	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		
Disciplinas			X	X	X	X	X	X	X	X	X	X			X	X	X	X	X	X	X	X	X	X		
Formulação da Rede								X	X	X	X	X	X	X	X											
Treino												X	X	X	X	X	X	X	X	X	X	X	X	X		
Resultado																					X	X	X	X	X	
Artigo 1																									X	X
Artigo 2																										
Tese																										

Tabela 7.1: Cronograma das atividades previstas para o primeiro biênio.

O estágio atual do projeto apresenta alguns resultados concernentes a dados sintéticos de uma rede neuronal e dois classificadores. Esses dados preliminares

fazem parte da investigação inicial de como resolver o problema proposto na Tese de Doutorado. Estes resultados foram utilizados para publicar um resumo expandido no Congresso internacional de Copenhague promovido pela EAGE. Este trabalho encontra-se em análise pela comissão do congresso.

Possuo dados para enviar mais um trabalho para um congresso específico de redes Neuronais Artificiais, tema principal da Tese.

Etapa	Meses																								
	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	
Pesquisa na Literatura	X	X	X	X	X	X																			
Disciplinas																									
Formulação da Rede																									
Treino																									
Resultado		X	X	X	X	X	X	X																	
Artigo 1		X	X	X																					
Artigo 2					X	X	X	X	X																
Tese																	X	X	X	X	X	X	X		

Tabela 7.2: Cronograma das atividades previstas para o segundo biênio.

O projeto encontra-se na etapa de treinamento da rede, contudo já apresenta alguns resultados preliminares. Segundo a minha avaliação o projeto encontra-se dentro do cronograma previsto inicialmente.

# Referências Bibliográficas

- ADIBIFARD, M., TABATABAEI-NEJAD, S., KHODAPANAH, E., 2014, “Artificial Neural Network (ANN) to estimate reservoir parameters in Naturally Fractured Reservoirs using well test data”, *Journal of Petroleum Science and Engineering*, v. 122, pp. 585–594. ISSN: 09204105. doi: 10.1016/j.petrol.2014.08.007. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S0920410514002563>>.
- ARTERO, A. O., 2008, *Inteligência Artificial Teórica e Prática*. 1st ed. São Paulo, Livraria da Física. ISBN: 9788578610296. Disponível em: <[www.livrariadafisica.com.br](http://www.livrariadafisica.com.br)>.
- ARTUR, P. C., SOARES, P. C., 2008, “Paleoestruturas e petróleo na Bacia do Paraná, Brasil”, *Brazilian Journal of Geology*, v. 32, n. 4, pp. 433–448.
- BENAOUDA, D., WADGE, G., WHITMARSH, R. B., et al., 1999, “Inferring the lithology of borehole rocks by applying neural network classifiers to down-hole logs: An example from the Ocean Drilling Program”, *Geophysical Journal International*, v. 136, n. 2, pp. 477–491. ISSN: 0956540X. doi: 10.1046/j.1365-246X.1999.00746.x.
- BIZZI, A. L., SCHOBENHAUS, C., VIDOTTI, R. M., et al., 2003, *Geologia, tectônica e recursos minerais do Brasil: texto, mapas e SIG*. CPRM - Serviço Geológico do Brasil.
- BORGHI, L., 2002, “A Bacia do Paraná”, *Anuário do Instituto de Geociências - IGEO, Departamento de Geologia*.
- CATÉ, A., PEROZZI, L., GLOUGUEN, E., et al., 2017, “Machine learning as a tool for geologists”, *The Leading Edge*, v. 36, n. 6 (mar.), pp. 215–219.
- CHAKRAVARTHY, S., CHUNDURU, R., FANINI, O., et al., 1999, “Detection of layer boundaries from array induction tool responses using neural networks”, *69th Ann. Internat. Mtg*, pp. 140–143. doi: 10.1190/1.1820779.

- CRISTINA LOPES QUINTAS, M., SILVIA MARIA MANTOVANI, M., VICTOR ZALÁN, P., 1999, “Contribuição Ao Estudo Da Evolução Mecânica Da Bacia Do Paraná”, *Revista Brasileira de Geociências*, v. 29, n. 2, pp. 217–226.
- EIRAS, J. F., 1996, “Influência da tectônica do Arco de Carauari na sedimentação fanerozóica da Bacia do Solimões, norte do Brasil. In Congresso Brasileiro de Geologia, 39. Salvador, Bahia.” v. 1, pp. 52–53.
- FELDMAN, J. A., FANTY, M. A., GODDARD, N. H., 1988, “Computing With Structured Neural Networks.” *Computer*, v. 21, n. 3, pp. 91–103. ISSN: 00189162. doi: 10.1109/2.34.
- FRANCA, ALMERIO & POTTER, E., 1991. “Reservoir Potential of glacial deposits of the Itarare Group.pdf” . .
- FREUND, Y., MASON, L., 1999, “The alternating decision tree learning algorithm”, *International Conference on Machine Learning*, v. 99, pp. 124–133. ISSN: 14602431. doi: 10.1093/jxb/ern164.
- HAGAN, M. T., DEMUTH, H. B., BEALE, M. H., et al., 1996, “Neural Network Design”, p. 1012. Disponível em: <<http://books.google.ru/books?id=bUNJAAAACAAJ>>.
- HALL, P., DEAN, J., KABUL, I. K., et al., 2014, “An Overview of Machine Learning with SAS ® Enterprise Miner™”, , n. Rosenblatt 1958, pp. 1–24.
- HAWKESWORTH, C., GALLAGHER, K., KIRSTEIN, L., et al., 2000, “Tectonic controls on magmatism associated with continental break-up an example from the Paraná Etendeka Province”, *Earth and Planetary Science Letters*, v. 179 (jun.), pp. 335–349.
- KANAL, L. N., 2001, “Perceptrons”, *Encyclopedia of Computer Science*, pp. 11215–11218. Disponível em: <[http://www.sciencedirect.com/science/article/B7MRM-4MT09VJ-2HC/2/11e1b0f217b506994f9009281762b4b5\\$!delimiter" data-bbox="228 775 800 875">B7MRM-4MT09VJ-2HC/2/11e1b0f217b506994f9009281762b4b5\\$!delimiter" data-bbox="228 775 800 875">026E30F\\$nhttp://www.sciencedirect.com/science?{\\_}ob=ArticleURL&{\\_}udi=B7MRM-4MT09VJ-2HC&{\\_}rdoc=8&{\\_}hierId=151000304&{\\_}refWorkId=21&{\\_}explode=151000302,151000304&{\\_}fmt=high&{\\_}o>](http://www.sciencedirect.com/science/article/B7MRM-4MT09VJ-2HC/2/11e1b0f217b506994f9009281762b4b5$!delimiter)

- KONATÉ, A. A., PAN, H., KHAN, N., et al., 2014, “Prediction of porosity in crystalline rocks using artificial neural networks: An example from the Chinese Continental Scientific Drilling Main hole”, *Studia Geophysica et Geodaetica*, v. 59, n. 1, pp. 113–136. ISSN: 00393169. doi: 10.1007/s11200-013-0993-5.
- KROGH, A., 2008, “What are artificial neural networks?” *Nature biotechnology*, v. 26, n. 2, pp. 195–197. ISSN: 1546-1696. doi: 10.1038/nbt1386. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/18259176>>.
- KUMAR, R., AGGARWAL, R., SHARMA, J., 2015, “Comparison of regression and artificial neural network models for estimation of global solar radiations”, *Renewable and Sustainable Energy Reviews*, v. 52, pp. 1294–1299. ISSN: 13640321. doi: 10.1016/j.rser.2015.08.021. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1364032115008679>>.
- LEVY, S., 1997, “The Computer”, *Newsweek*, v. 130, n. 22, pp. 28. Disponível em: <<file:///D:/Dropbox/Whitfield/History/SemesterTwo/ChangingTimes/Research/EBSCOhost2.htm>>.
- MACKAY, D. J. C., 2005, *Information Theory, Inference, and Learning Algorithms* David J.C. MacKay, v. 100. ISBN: 9780521642989. doi: 10.1198/jasa.2005.s54. Disponível em: <<http://pubs.amstat.org/doi/abs/10.1198/jasa.2005.s54>> <<http://www.cambridge.org/0521642981>>.
- MAO, J., 1996, “Why artificial neural networks?” *Communications*, v. 29, pp. 31–44. ISSN: 00189162. doi: 10.1109/2.485891. Disponível em: <[http://ieeexplore.ieee.org/xpls/abs{\\_}all.jsp?arnumber=485891](http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=485891)>.
- MCCULLOCH, W. S., PITTS, W., 1943, “A logical calculus of the ideas immanent in nervous activity”, *The Bulletin of Mathematical Biophysics*, v. 5, n. 4, pp. 115–133. ISSN: 00074985. doi: 10.1007/BF02478259.
- MICHEL, M. D., DEZA, E., 2016, *Encyclopedia of Distances*. ISBN: 978-3-662-52844-0. doi: 10.1007/978-3-662-52844-0.
- MICHIE, E. D., SPIEGELHALTER, D. J., TAYLOR, C. C., 1994, “Machine Learning , Neural and Statistical Classification”, *Technometrics*, v. 37, n. 4, pp. 459. ISSN: 00401706. doi: 10.2307/1269742. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.94.3615>>.

- MILANI, E. J., RAMOS, V. A., 1998a, “Paleozoic orogenies in southwestern Gondwana and the subsidence cycles of the Parana Basin\Orogenias paleozoicas no domínio sul-occidental do Gondwana e os ciclos de subsidência da bacia do Paraná”, *Revista Brasileira de Geociencias*, v. 28, n. 4, pp. 473–484.
- MILANI, E. J., BRANDÃO, J., ZALÁN, P. V., et al., 2000, “Petróleo na margem continental Brasileira: Geologia, exploração, resultados e perspectivas”, *Revista Brasileira de Geofísica*, v. 18, n. 3, pp. 351–396. ISSN: 0102261X. doi: 10.1590/S0102-261X2000000300012.
- MILANI, E. J., RAMOS, V. A., 1998b, “Orogenias paleozóicas no domínio sul-occidental do Gondwana e os ciclos de subsidência da Bacia do Paraná”, *Brazilian Journal of Geology*, v. 28, n. 4, pp. 473–484.
- MILANI, E. J., ZALAN, P. V., 1999, “An outline of the geology and petroleum systems of the Paleozoic interior basins of South America”, *Milani1999*, v. 22, pp. 199–205. Disponível em: <<http://www.episodes.co.in/www/backissues/223/199-205Milani.pdf>>.
- MILANI, E., SPADINI, A., TERRA, G., et al., 2007, *Boletim de geociências da Petrobras*, v. v. 15. Cenpes.
- MINSKY, M., PAPERT, S., 1969, *Perceptrons*. doi: 10.1016/j.neucom.2015.05.138.
- MISRA, J., SAHA, I., 2010, “Artificial neural networks in hardware: A survey of two decades of progress”, *Neurocomputing*, v. 74, n. 1-3, pp. 239–255. ISSN: 09252312. doi: 10.1016/j.neucom.2010.03.021. Disponível em: <<http://dx.doi.org/10.1016/j.neucom.2010.03.021>>.
- MOHRIAK, W., SZATMARI, P., ANJOS, S., 2008, *Sal: Geologia e Tectônica. Exemplos nas Bacias Brasileiras*. 1 ed. São Paulo, SP., Beca. ISBN: 978-85-87256-49-2.
- NEDJAH, N., DA SILVA, F. P., DE SÁ, A. O., et al., 2016, “A massively parallel pipelined reconfigurable design for M-PLN based neural networks for efficient image classification”, *Neurocomputing*, v. 183, pp. 39–55. ISSN: 18728286. doi: 10.1016/j.neucom.2015.05.138.
- POULTON, M. M., 2002, “Neural networks as an intelligence amplification tool: A review of applications”, *Geophysics*, v. 67, n. 3, pp. 979. ISSN: 0016-8033. doi: 10.1190/1.1484539.

- ROSENBLATT, F., 1962, “Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms.” *Archives of General Psychiatry*, v. 7, pp. 218–219. ISSN: 0003-990X. doi: 10.1001/archpsyc.1962.01720030064010.
- SALJOOGHI, B. S., HEZARKHANI, A., 2014, “Comparison of WAVENET and ANN for predicting the porosity obtained from well log data”, *Journal of Petroleum Science and Engineering*, v. 123, pp. 172–182. ISSN: 09204105. doi: 10.1016/j.petrol.2014.08.025.
- SCHERER, C. M. S., LAVINA, E. L. C., 2006, “Stratigraphic evolution of a fluvial-eolian succession: The example of the Upper Jurassic-Lower Cretaceous Guara and Botucatu formations, Parana Basin, Southernmost Brazil”, *Gondwana Research*, v. 9, n. 4, pp. 475–484. ISSN: 1342937X. doi: 10.1016/j.gr.2005.12.002.
- TELFORD, W. M., SHERIFF, R. E., 1993, *Applied Geophysics*. Cambridge University Press.
- VAIL, P. R., MITCHUM, R. M., THOMPSON, S., 1977, *Seismic stratigraphy and global changes of sea level*. Seismic stratigraphy: applications to hydrocarbon exploration. APPG.
- YAN, Z., XIAODONG, Z., JIAOTONG, L., et al., 2014, “Lithology identification research based on self-organizing map of data mining method”, .
- ZALAN, P. V., WOLF, S., 1987, “Tectônica e Sedimentação da Bacia do Paraná”. In: *Simpósio sul-brasileiro de geologia, SBG, 3, Atas, Curitiba-PR.*, v. 1, pp. 441–477.
- ZALAN, P. V., 2007, “Evolução Fanerozóica das Bacias Sedimentares Brasileiras”. In: *Geologia da Plataforma Sul-Americana*, Petrobras, cap. 23, pp. 595–613, Rio de Janeiro.
- ZHANG, L., POULTON, M., ZHANG, Z., et al., 1999, “Fast forward modeling simulation of resistivity well logs using neural network”, *69th Ann. Internat. Mtg*, pp. 124–127. doi: 10.1190/1.1820734.

## Introduction

Machine learning field approaches the creation of computer programm's that have the capability of automatically improve themselves through experience (Michie et al., 1994; Levy, 1997; MacKay, 2005). Classification techniques such as Euclidean and Mahalanobis similarity measurements are considered classical machine learning methods. Those similarity measurements are referred to as distance attributes (Marie and Deza, 2016). Euclidean classifiers include a calculation of a centroid on the space of attributes. While the Mahalanobis takes into consideration the shape of attributes space. Both techniques are capable of making identification of geological cyclicity in data.

A Self Organizing Map (SOM) is inspired by neural cortex (Kohonen, 1989). A SOM algorithm is based on a network (Haykin, 2001). This geometric arrangement is an oriented graph, whose vertices are the fundamental units know as artificial neurons and the edges are weights governing the interactions among neurons. Those artificial neurons change their weights as interactions go on.

This work aims to define a comparison between a Kohonen SOM, an euclidean and a mahalanobean classifiers. This comparison uses two well log data from a synthetic syneclyses sedimentary basin type. It is remarkable that the Mahalanobis classifier produced a higher error when compared to the Euclidean classifier and the SOM. The SOM presented better results for the two synthetic examples, with an error of 0.7% for the first well and 1.5% for the second. In contrast, Mahalanobis and Euclidean classifiers presented an error of 18.3% and 1.7% respectively for the first well and 11.3% and 6% for the second.

## Methodology

In a general overview, the methodology adopted in this work is divided into three main parts. The first generates a synthetic syneclyses sedimentary basin in which three synthetic wells are drilled (see Fig. 1). The second part uses well log T1 to train the Kohonen SOM and obtain an optimal distribution of weights. Additionally, the same well is again used to store T1 log data into arrays. The last is to use the three techniques and compare the classified patterns for wells C1 and C2.

### Synthetic Sedimentary Basin

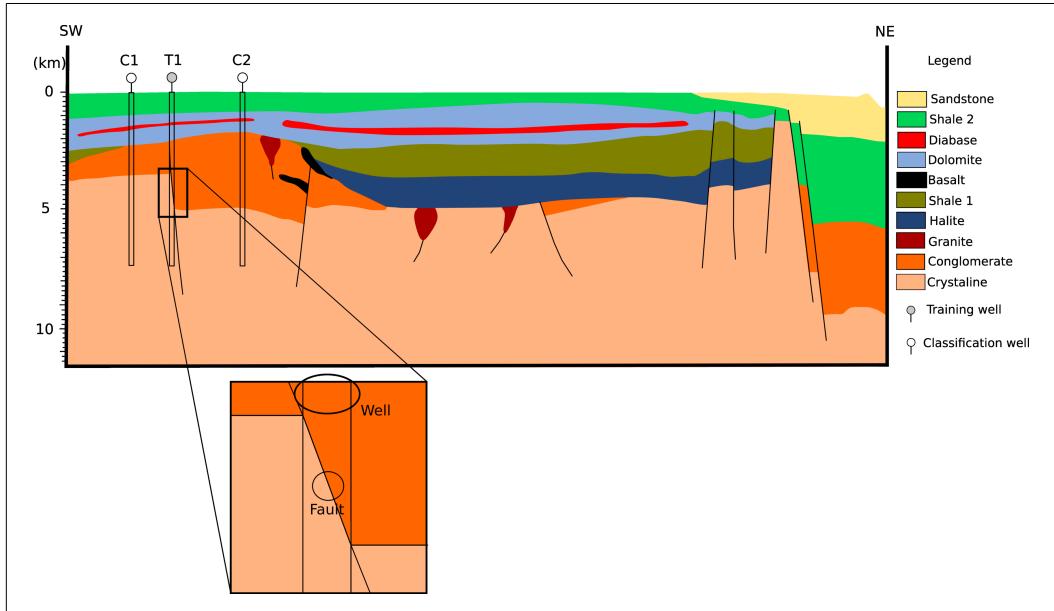
The proposed model for the machine learning tests was based on a schematic geological model proposed by Mohriak et al. (2008) for the Solimões Sedimentary Basin, North part of Brazil. This modelling reproduces structures such as Horts, Grabens, normal and reverse faults. Fig. 1 shows the model with a zoom box highlighting the non-parallel contact of two different lithotypes, where three wells were sampled. Four physical data properties were considered: density, gamma-ray, resistivity and velocity (see Tab. 1). The sample rate for the well data is 0.01 observation/meter with contamination of 5% gaussian noise.

Rock	Density ( $g/cm^3$ )	Gamma-ray ( $Ci/g$ )	Resistivity ( $\Omega/m$ )	Velocity ( $Km/s$ )
Conglomerate	2.30	100.0	6000	2
Shale	2.55	100.0	1000	3
Dolomite	2.72	8.30	$3.5 \times 10^3$	6
Diabase	2.91	30.0	$15 \times 10^7$	5.5
Crystalline	2.80	0.7	$1.3 \times 10^6$	5

**Table 1** Physical properties.

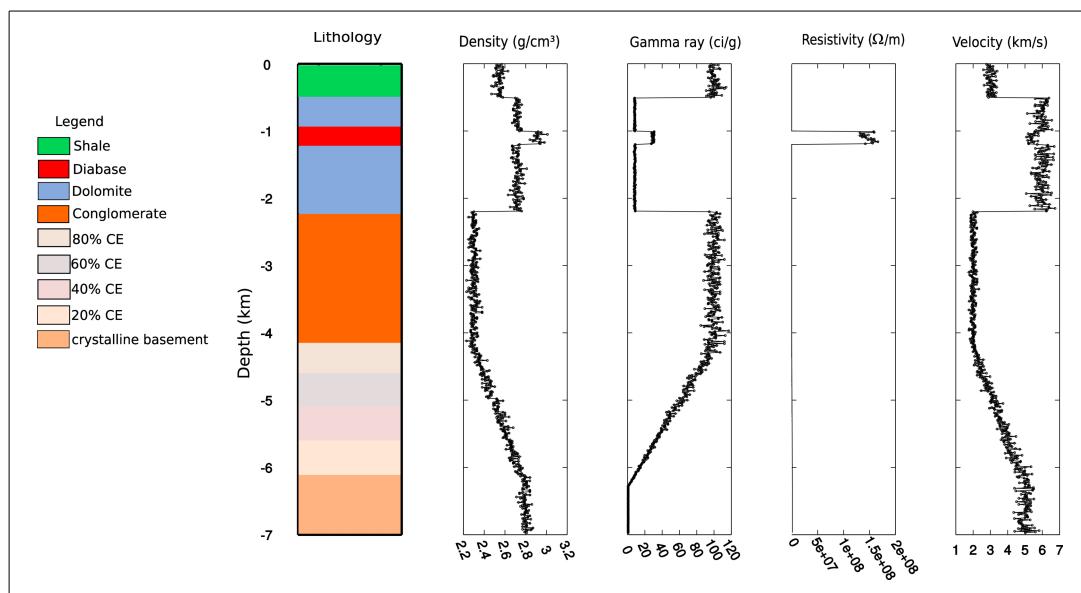
### Training and Similarities

SOM are machine learning types composed by oriented graphs that are distributed on a hyperplane inside a hyperspace of features. Features are the physical properties that are correlated with a specific type of



**Figure 1** Synthetic Sedimentary Basin by Mohriak et al. (2008) T1, C1 and C2 are training and classifying wells respectively.

rock. The identification process is based on redundancy of patterns. A toroid geometry is adopted here for the SOM with 400 neurons. Vector  $\mathbf{X}$  is composed of physical properties from the training well T1 with dimension  $n$ , where  $n$  is the number of data.  $\mathbf{X}$  is related to a neuron with a  $w_{i,j}$  weight matrix, as follows:



**Figure 2** Synthetic training well T1. The normal fault was created by four divisions on the range of the normal fault, decreasing the amount of conglomerate in comparison to cryshalline basement. That behavior simulates a special kind of logging signature.

$$d(t) = \sqrt{\sum_{i=1}^n [x_i(t) - w_{i,j}(t)]^2} \quad (j = 1, \dots, m), \quad (1)$$

where  $t$  is the number of iterations,  $x(t)$  is an element of  $\mathbf{X}$  and  $d(t)$  is the metric. The lowest value of  $d(t)$  defines the best neuron for a specific attribute.

The Euclidean classifier is a statistical non-parametric classifier that uses the same input vector  $\mathbf{X}$  for its training process. A mean vector  $\bar{\mathbf{X}}_i$  for each property is calculated and stored in a set of training  $i$ . Then the Euclidian distance for the  $i$ -th set ( $Ed_i$ ) is computed as the following:

$$Ed_i = \|\mathbf{X} - \bar{\mathbf{X}}_i\|^{\frac{1}{2}}, \quad (2)$$

where  $\mathbf{X}$  is the set of physical properties to be classified.

The Mahalanobis Classifier computes the mean vector  $\bar{\mathbf{X}}_i$  and the covariance matrix  $\mathbf{C}_i$  for each ensemble. The Mahalanobis distance ( $Md_i$ ) is computed as:

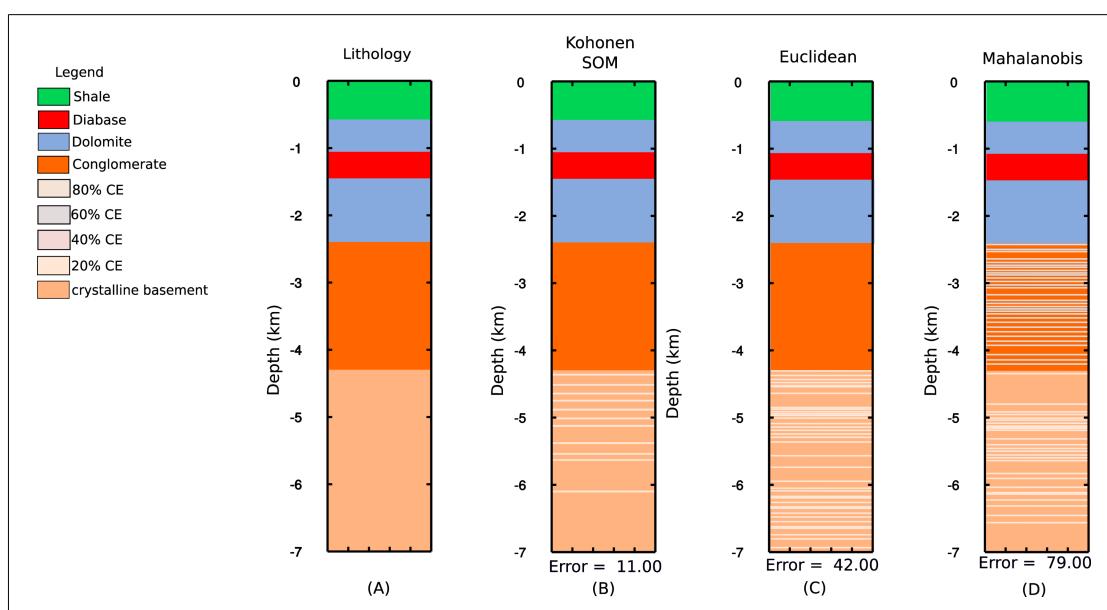
$$Md_i = [(\mathbf{X} - \bar{\mathbf{X}}_i)^T \mathbf{C}_i^{-1} (\mathbf{X} - \bar{\mathbf{X}}_i)]^{\frac{1}{2}}. \quad (3)$$

The covariance matrix is defined as:

$$\mathbf{C}_i = \frac{1}{n_i - 1} \sum_{X \in \omega_i} (\mathbf{X} - \bar{\mathbf{X}}_i)(\mathbf{X} - \bar{\mathbf{X}}_i)^T \quad (4)$$

## Results

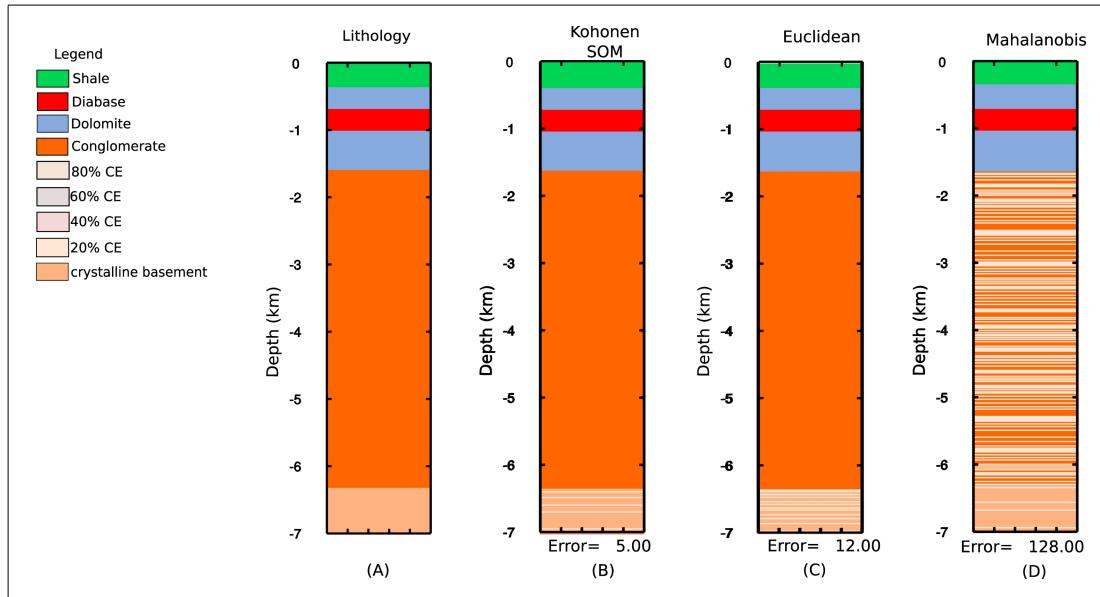
Fig. 3 (A) shows the original well. Fig. 3 (B), (C) and (D) present the final classification for SOM, Euclidean and Mahalanobis. All errors were concentrated on a single lithotype, the crystalline rock. Those errors indicate 11 swaping between crystalline rock and 20%CE rock type for the SOM classificator.



**Figure 3** Comparison between the classifiers and SOM for C1 well data.

Euclidean classifier results on 42 errors. Those errors present the same pattern of SOM classifiers. Mahalanobis classifier shows 79 errors. It misjudges the crystalline rock data with the 20%CE rock and the conglomerate with the 60%CE rock.

Fig. 4 (A) shows the synthetic well. Fig. 4 (B), (C) and (D) present the comparison among the three methodologies. In a overall perspective, SOM shows better results. Again for the three methods, the major misleading occurs in the classification of crystalline basement.



**Figure 4** Comparison between the classifiers and SOM for C2 well data.

## Conclusions

Two synthetic tests of well logging were performed to understand the behavior of three different machine learning elements: Kohonen SOM, Euclidean and Mahalanobis classifiers. As expected, the SOM overperformed the classifiers due to a more detailed algorithm. On the other side, the computational requirements for SOM are more demanding than the classifiers, which indicates that the choice of the method depends on the number of data sets.

As perspectives, we are intended to apply these methods to more complex synthetic scenarios and also with real data acquired on Paraná Sedimentary Basin, South portion of Brazil.

## References

- Haykin, S. [2001] *Neural networks: principles and practice*. McMaster University, Hamilton, Ontario - Canada., 2 edn.
- Kohonen, T. [1989] Biological Cybernetics 9 1989. **425**, 139–145.
- Levy, S. [1997] The Computer. *Newsweek*, **130**(22), 28.
- MacKay, D.J.C. [2005] *Information Theory, Inference, and Learning Algorithms* David J.C. MacKay, 100.
- Marie, D.M. and Deza, E. [2016] *Encyclopedia of Distances*. Springer, Moscow State Pedagogical University, 4 edn.
- Michie, E.D., Spiegelhalter, D.J. and Taylor, C.C. [1994] Machine Learning , Neural and Statistical Classification. *Technometrics*, **37**(4), 459.
- Mohriak, W., Szatmari, P. and Anjos, S. [2008] *Salt: Geology and Tectonic. Exemples on Brazilian's Sedimentary Basins. (In Portuguese)*. Beca, São Paulo, SP., 1 edn.