Relatório Parcial: Ano 2

Isaac L. Santos Sacramento
Orientador: Mauro Roisenberg

October 17, 2016

Abstract

Neste relatório são apresentadas as atividades realizadas no segundo ano de projeto, mais precisamente no período de março a outubro de 2016. Neste período foram realizados estudos e experimentos relacionados à aplicação de *Deep Learning* para realizar simulação geoestatística multiponto. Estas atividades estão descritas nas seções a seguir.

1 Introdução

No período de março a outubro de 2016 foram exploradas duas novas frentes de trabalho relacionados ao estudo da incerteza. Inicialmente, foram estudados os algoritmos de simulação multiponto, por meio dos quais é possível gerar realizações de fenômenos geoestatísticos condicionadas a amostras reais disponíveis. A simulação multiponto utiliza imagens de treinamento como dados de entrada para definir a estrutura espacial de um processo [2] e difere dos métodos baseado em variograma, pois utilizam um conjunto de pontos na vizinhança do ponto simulado. As imagens de treinamento são descrições explícitas, 2D ou 3D, da continuidade espacial estudada e, a partir delas, podem ser extraídas estatísticas de múltiplos pontos, inclusive o variograma. Em particular, o algoritmo FILTERSIM despertou o interesse na utilização de técnicas de *Deep Learning* para realização de simulação geoestatística devido a sua similaridade com a definição de Redes Neurais Convolucionais. O algoritmo é descrito com maior detalhe na seção seguinte, juntamente com outros algoritmos de simulação geoestatística multiponto.

A etapa do trabalho está relacionada à utilização de técnicas de *Deep Learning* e seu potência de aplicação para solução de problemas relacionados a simulação geoestatística. Foram realizados estudos para a definição de um método capaz de realizar simulação multiponto utilizando o conceito de *Deep Learning*. As redes neurais convolucionais (RNC) estão

no centro das pesquisas relacionadas a aprendizagem de máquina. As RNC são muito populares nas áreas de classificação de imagens [4, 3], visão computacional [15], reconhecimento de faces [13], à segmentação de imagem e reconhecimento de ações em vídeos; Os fatores que conferem às RNC a importância atual são: (i) a eficiência no treinamento em GPUs modernas, (ii) a proposta das Unidades Lineares Retificadas (em tradução livre do termo em inglês Rectified Linear Unit - ReLU) as quais tornam a convergência mais rápida, e (iii) o acesso amplo aos dados para treinamento de modelos extensos (ImageNet)[4]. Semelhante aos algoritmos de simulação geoestatística multiponto, as redes neurais convolucionais utilizam imagens de treinamento como dados de entrada. Como visto anteriormente, as RNC são amplamente utilizadas para resolver problemas de classificação. Entretanto, visando realizar simulações geoestatísticas, experimentos são realizados com diferentes conjuntos de dados a fim de obter um modelo regressivo de redes convolucionais.

2 Revisão Bibliográfica

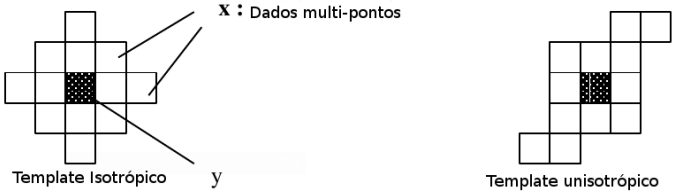
Os métodos tradicionais de simulação se apoiam na modelagem de estatísticas em dois pontos, geralmente as covariâncias e variogramas. Muitos fenômenos são complexos e inviabilizam a captura de seus padrões espaciais por meio de estatísticas de dois pontos. A simulação geoestatística multiponto é um método genérico e se baseia em três mudanças conceituais formalizadas por [2]. A primeira, afirma que conjuntos de dados podem não ser suficiente para inferir todas as características estatísticas que controlam o que se deseja modelar. A segunda é adotar uma estrutura estatística não-paramétrica para representar a heterogeneidade. A terceira mudança conceitual é avaliar a estatística de eventos de dados de múltiplos pontos. As estatísticas multipontos são expressas como funções densidades cumulativas para uma variável aleatória $Z(x)$ condicionadas a eventos de dados locais $d_n = Z(x_1), Z(x_2), ..., Z(x_n)$, isto é, os valores de Z nos nós vizinhos x_i de x, equação 1.

$$f(z, x, d_n) = Prob(Z(x) <= z|x) \tag{1}$$

Os diferentes algoritmos que compõem este método compartilham elementos comuns,

como o uso de um caminho de simulação (geralmente aleatório), a amostragem em distribuições de probabilidades locais e o uso de grades múltiplas [6]. [1] apresentam um modelo de simulação multiponto baseado em redes neurais que utilizam imagens de treinamento como fonte de aprendizado. Deste modo, as redes neurais podem ser ensinadas a coletar estatísticas de múltiplos pontos e utilizá-las para gerar modelos probabilísticos condicionados aos dados reais. Para realizar esta tarefa, as distribuições condicionais aprendidas pela rede, relacionam o valor em qualquer região da imagem aos dados da sua vizinhança, dentro de um *template* centrado na região simulada, semelhante à ilustração da figura 1. A rede

Figure 1: Definição de uma vizinhança.



Fonte: [1]

neural é treinada para determinar os valores de y em qualquer região, dados os valores da vizinhança x. Matematicamente, a rede neural modela a função densidade de probabilidade (FDP) local da equação 2, ou sua integral, a função de distribuição condicional cumulativa (FDCC).

$$f(y|x) = Pr(y < Y < y + dy|x) \tag{2}$$

A abordagem multiponto baseada em redes neurais [1], fortaleceu o interesse no estudo dos métodos multipontos. Em 2002, o algoritmo *SNESIM* [12] foi apresentado como uma alternativa para realizar simulações sequenciais livres de variogramas e capazes de lidar com a presença de padrões não-estacionários nas imagens de treinamento. Este algoritmo utiliza estrutura de árvore para armazenar a distribuição de probabilidade condicional calculada a partir da imagem de treinamento. Em 2011, [11] apresentaram o método *IMPALA*, que consiste na implementação do método multiponto original, com a substituição da árvore de busca por uma estrutura de lista. Esta medida permite reduzir a quantidade de memória RAM utilizada pelo algoritmo e, consequentemente, utilizar diferentes listas para guardar dados adicionais igualmente utilizados durante a simulação. Em 2010 o algoritmo *simu-*

lated annealing foi implementado com intuito de gerar realizações estocásticas de variáveis categóricas por reprodução de estatísticas multipontos [9]. A imagem de treinamento é utilizada para determinar as frequências de ocorrência das configurações de cada nó. Estas frequências são usadas como estatística alvo que deve combinar com as imagens estocásticas geradas com o algoritmo.

Os algoritmos, *SNESIM* e *IMPALA*, se baseiam no armazenamento dos eventos de dados encontrados na imagem de treinamento, são restritos à simulação de variáveis categóricas e necessitam de um modelo para a vizinhança do ponto simulado. Uma nova abordagem para SMP procede com amostragens diretamente sobre a imagem de treinamento para um determinado evento, de modo que o uso de um banco de dados de evento se torna dispensável [7] [8]. O método de amostragem direta permite estender a aplicação de geoestatística multiponto para variáveis contínuas e categóricas e se destaca por respeitar a distribuição de probabilidade condicional sem realizar a sua computação. Na utilização de múltiplas variáveis, uma função d é escolhida apropriadamente para variáveis categóricas ou contínuas. É possível utilizar a função de distância ($d(.)$) para controlar as proporções globais e locais, ou impor uma média local. Para tanto, é adicionado um fator de erro (E_p) à fdp, que quantifica a diferença entre a probabilidade alvo $L(x)$ e a proporção atual $P(x)$ na grade SG parcialmente simulada, na forma da equação 3.

$$E_p \, |P(x) - L(x)| \tag{3}$$

Uma observação comum na comparação entre a imagem de treinamento e as realizações geradas é a existência de padrões exatamente reproduzidos da imagem de treinamento para a grade de simulação. Este fenômeno é uma consequência de coerência de textura e o tamanho limitado da imagem de treinamento. Algoritmos de simulação baseados em fragmentos ou blocos (*patch*) tendem a agrupar valores de simulação que estejam próximos uns dos outros na imagem de treinamento. O algoritmo de simulação baseada em filtro, *FILTERSIM*, foi concebido para superar a limitação do *SNESIM* em funcionar apenas para variáveis categóricas. O *FILTERSIM* agrupa todos os padrões obtidos da imagem de treinamento dentro de um conjunto de classes de padrões e, em cada local de simulação,

identifica a classe de padrão que mais se assemelha ao evento de dado condicionante. Em seguida, um padrão é amostrado dentro do protótipo de padrões e copiado para a grade de simulação [16]. Na versão original do algoritmo, a escolha da classe de padrão é baseada na distância de pixel entre cada classe de padrão e o evento de dado condicionante, o que torna o algoritmo custoso computacionalmente para simulações 3D. Entretanto, uma nova abordagem é adotada, na qual o cálculo da distância é substituído por comparação de pontuação de filtro, ou seja, a diferença da pontuação de filtro entre o padrão condicionante e cada protótipo de padrão [14]. O cálculo baseado em pontuação permite reduzir o custo computacional por conta da redução dimensional dos dados.

Este algoritmo é particularmente interessante, pois ele aborda a simulação multiponto de modo similar ao princípio da convolução utilizado pelas redes neurais convolucionais para extrair padrões durante o processo de treinamento. Neste sentido, pode é relevante investigar a aplicação destas redes para realizar a simulação geoestatística multiponto, bem como estimar propriedades petrofísicas.

3 Convolutional Neural Network Toolbox

Os primeiros experimentos realizados tratam da utilização da textittoolbox para redes neurais convolucionais, desenvolvida para a plataforma MATLAB. A textitToolbox implementa um modelo convolucional para classificação de dígitos numéricos escritos á mão e utiliza o banco de dados MNIST [5], que contém 60 mil exemplos para treinamento, validação e teste do modelo. Devido a particularidade de ter sido desenvolvida para um conjunto de dados específico, a estrutura da rede necessita ser modificada sempre que um novo conjunto de dados é utilizado como entrada para a rede. Duas alterações estão em andamento, a primeira para utilização de dados de log de poços como entrada para o treinamento da rede e a segunda, realizada na camada de saída da rede, para torná-la regressiva e com a saída tendo a mesma dimensão da imagem de entrada.

Os pesos das camadas convolucionais, iniciados aleatoreamente incialmente e ajustados durante treinamento, foram substituídos por filtros de Gabor e filtros de média, gradiente e curvatura, definidos no algoritmo FILTERSIM [14]. As modificações causaram uma mu-

dança no desempenho da rede para o problema de classificação, mas dependem de uma análise mais apurada para chegar e a aplicação no problema de regressão para chegar a uma conclusão mais precisa.

4 Caffe Framework

O segundo experimento em andamento está relacionado com a utilização do *Framework* Caffe. Este *Framework* implementa os algoritmos de treinamento de Deep Learning, inclusive o modelo convolucional. Para tanto, os dados de imagem são carregados para um banco de dados LMDB por meio de *scripts* do próprio Caffe. A implementação do modelo de rede, assim como os parâmetros de treinamento; tais como função de ativação das camadas, tamanho do *batch* de dados, taxa de decaimento, número de iterações, momentum, entre outros; são definidos em arquivo texto com extensão ***.prottxt**, que será lido e interpretado pelo *Framework* para gerar um modelo de rede neural convolucional. O framework está em teste e, em caso de sucesso na elaboração de um modelo regressivo, será utilizado com dados de imagens de logs de poços para a simulação de propriedades petrofísicas.

5 Conclusões

A simulação multiponto gera realizações que reproduzam padrões estatísticos inferidos a partir de alguma fonte, usualmente uma imagem de treinamento. Como ferramenta de modelagem, os algoritmos de aprendizagem de máquina são universais, adaptativos, não-lineares, robustos e eficientes. Eles podem alcançar soluções aceitáveis para problemas de classificação, regressão e modelagem de densidade de probabilidade em espaço de alta dimensão e com características espacialmente referenciados. Dentre os raros trabalhos encontrados estão o uso de algoritmos genéticos na simulação de variáveis categóricas para reproduzir estatísticas multipontos [10]. Este algoritmo requer alto desempenho computacional e, portanto, depende da disponibilidade de computadores com múltiplos núcleos, bem como unidades de processamentos gráficos, características inerentes aos métodos de algoritmos genéticos. A simulação multiponto com redes neurais apresentada por Jef Caers [1] explora uma solução neural para a simulação pixel a pixel e, embora haja diversas citações

a este trabalho, nenhuma delas está relacionada com a melhoria, expansão ou aplicação do método proposto. Este fato se dá, possivelmente, por conta do interesse no desenvolvimento de novos algoritmos.Com base no levantamento literário, se observa a possibilidade de explorar a simulação multiponto por meio da implementação baseada em aprendizagem de máquina. Os experimentos de simulação multiponto com os métodos de *Deep Learning* estão em andamento e parecem promissores para este fim, entretanto, ainda não apresentaram resultados conclusivos.

References

[1] Jef Caers and Andre Journel. Stochastic reservoir simulation using neural networks trained on outcrop data. *SPE Annual Technical Conference and Exhibition*, 1998.

[2] Felipe B. Guardiano and R. Mohan Srivastava. *Geostatistics Tróia 92: Volume 1*. Springer Netherlands, Dordrecht, 1993.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition*, pages 346–361. Springer International Publishing, Cham, 2014.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, page 2012.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 28–2324, 1998.

[6] Gregoire Mariethoz and Jef Caers. *Multiple-Point Geostatistics Algorithms*. John Wiley & Sons, Ltd, 2014.

[7] Gregoire Mariethoz, Philippe Renard, and Julien Straubhaar. The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research*, 46(11):n/a–n/a.

[8] Eef Meerschman, Guillaume Pirot, Gregoire Mariethoz, Julien Straubhaar, Marc Van Meirvenne, and Philippe Renard. A practical guide to performing multiple-point statistical simulations with the direct sampling algorithm. *Computers and Geosciences*, 52:307–324, 2013.

[9] Oscar Peredo and JuliÃ¡n M. Ortiz. Parallel implementation of simulated annealing to reproduce multiple-point statistics. *Computers & Geosciences*, 37(8):1110 – 1121, 2011.

[10] Oscar Peredo and Julián M. Ortiz. *Geostatistics Oslo 2012*, chapter Multiple-Point Geostatistical Simulation Based on Genetic Algorithms Implemented in a Shared-Memory Supercomputer, pages 103–114. Springer Netherlands, Dordrecht, 2012.

[11] Julien Straubhaar, Philippe Renard, Grégoire Mariethoz, Roland Froidevaux, and Olivier Besson. An improved parallel multiple-point algorithm using a list approach. *Mathematical Geosciences*, 43(3):305–328, 2011.

[12] Sebastien Strebelle. Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, 34(1):1–21, 2002.

[13] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In Z. Ghahramani, M. Welling, C. Cortes, N.d. Lawrence, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1988–1996. Curran Associates, Inc., 2014.

[14] Jianbing Wu, Tuanfeng Zhang, and André Journel. Fast filtersim simulation with score-based distance. *Mathematical Geosciences*, 40(7):773–788, 2008.

[15] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. *Part-Based R-CNNs for Fine-Grained Category Detection*, pages 834–849. Springer International Publishing, Cham, 2014.

[16] Tuanfeng Zhang, Paul Switzer, and Andre Journel. Filter-based classification of training image patterns for spatial simulation. *Mathematical Geology*, 38(1):63–80, 2006.