ISSN: 2446-8339

Avaliação do efeito da variação de dimensionalidade na seleção de realizações geoestatísticas representativas para quantificação de risco, usando o método de Análise de Componente Principal

Henrique Hungari Rodrigues [1]
Luís Otávio Mendes da Silva [2]
Susana Margarida da Graça Santos [3]
Denis José Schiozer [4]

RESUMO

A etapa de seleção de modelos representativos (MRs) para tomada de decisão sob incerteza tem, muitas vezes, um elevado custo computacional gasto em simulações de escoamento (Schiozer et al., 2019). Como forma de reduzir este custo, Mahjour et al. (2020) simplificaram as 300.000 dimensões das realizações geoestatísticas do modelo benchmark UNISIM-II-D em duas, utilizando redução de dimensionalidade. Contudo, tal simplificação tem como consequência a perda da variabilidade do conjunto de dados. Assim, esse trabalho utiliza Análise de Componente Principal (PCA, na sigla em inglês) para redução de dimensionalidade, variando o número de dimensões geradas para avaliar a quantidade de informação capturada no sistema simplificado, e buscar a melhor configuração do fluxo de trabalho para quantificação de risco. Foi observado que, para o caso estudado, as dimensões geradas pela PCA capturam pouca variabilidade e de forma heterogenia em relação às propriedades que as dimensões representam, como porosidade. Dessa forma, um sistema simplificado com poucas dimensões, além de pouca informação, fica enviesado. Em relação à quantificação de risco, independentemente do número de MRs e técnicas de clusterização, o aumento do número de dimensões geradas não só não favoreceu os resultados como aumentou os erros relacionados à representação do risco. Este fenômeno é explicado pela literatura como a "maldição da dimensionalidade". Recomenda-se a aplicação do fluxo de trabalho usando poucas dimensões (entre 2 e 4), Kmeans como método de clusterização para seleção de MRs e o maior número de MRs possível.

Palavras-chave: Análise de Componente Principal; Modelos Representativos; UNISIM-II-D; Quantificação de Risco.

INTRODUÇÃO

O interesse econômico por petróleo faz com que companhias e governos busquem maneiras de maximizar a produção de reservatórios. A produção leva décadas, envolve infraestruturas de alto custo e incertezas de reservatório, econômicas e operacionais. Para determinar uma estratégia de produção no começo da operação, incertezas são complicadores

[1] Graduado do Curso de Eng. Mecânica da Unicamp, henrique.hungari@gmail.com
[2] Doutor pelo Curso de Eng. Mecânica da Unicamp, luis.otavio@cepetro.unicamp.br
[3] Co-orientadora: Doutora de Ciências do Petróleo da Unicamp, sgsantos@cepetro.unicamp.br
[4] Professor orientador: Doutor de Eng. Mecânica da Unicamp, denis@cepetro.unicamp.br

 ISSN: 2446-8339

que devem ser mitigados para garantir a escolha de uma estratégia confiável e eficiente. Schiozer et al. (2019) propuseram um fluxo de trabalho que reduz essas dificuldades, através de uma metodologia em malha fechada para suportar a tomada de decisão. Para isso, considera modelos de simulação de reservatório, análise de risco, assimilação de dados, redução de incertezas, modelos representativos (MRs) e seleção de estratégia de produção sob incerteza.

O fluxo de trabalho é muito eficaz, mas um dos obstáculos é o alto custo computacional relacionado às centenas de simulações de escoamento para geração de indicadores de produção. Para reduzir este custo, Mahjour et al. (2020) assumem que realizações geoestatísticas semelhantes teriam simulações de escoamento semelhantes e selecionam realizações representativas usando métodos de clusterização, que não requerem simulação de fluxo. Dessa forma, MRs são selecionados a partir de realizações geoestatísticas representativas, reduzindo o número de simulações necessárias.

Considerando um reservatório de petróleo representado em um modelo com milhares de blocos de características únicas, pode-se interpretá-lo como um sistema cartesiano onde cada propriedade de cada bloco é uma dimensão cartesiana. A seleção de realizações geoestatísticas usa clusterização baseada em distância, mas uma vez que modelos geoestatísticos contêm uma grande quantidade de dimensões, devem ser considerados como um conjunto de dados de alta dimensionalidade. Isso implica que pode existir uma grande quantidade de dados redundantes ou irrelevantes, aumentando o erro de medição de distância e prejudicando a clusterização, conforme explica a “maldição da dimensionalidade” por Bellman (1957). Para mitigar tal dificuldade, uma etapa de pré-processamento gera sistemas de baixa dimensionalidade usando métodos de redução de dimensionalidade, a partir de combinações lineares das dimensões. Mahjour et al. (2020) utiliza *Multidimensional Scaling* com redução para 2 dimensões e Kmedoids para seleção de MRs, deixando áreas em aberto para as pesquisas futuras. Assim, os objetivos desse trabalho são avaliar o efeito do número de dimensões sobre a quantidade de informação (variância do sistema original) capturada pelo sistema reduzido, e buscar a melhor configuração do fluxo de trabalho para quantificação de risco, variando número de dimensões, técnica de clusterização e número de MRs.

METODOLOGIA

Para cumprir as avaliações, o presente trabalho usa o método de Análise de Componente Principal (PCA, na sigla em inglês) para redução de dimensionalidade. É avaliado (a)

ISSN: 2446-8339

variabilidade capturada por cada nova dimensão do sistema e a composição, em propriedades petrofísicas, das novas dimensões mais relevantes; (b) a quantificação de risco ao variar o método de clusterização e número de novas dimensões; (c) a quantificação de risco ao variar o número de MRs selecionados e número de novas dimensões.

Para a execução do item (a), usando Python v3.6.9, os dados de realizações geoestatísticas são importados e normalizados (tornando a média nula e desvio-padrão unitário das amostras em cada uma das dimensões), de forma que diferença de escala entre as propriedades não interfira na importância das mesmas.

Na sequência, a PCA é executada, através da biblioteca Scikit-learn, gerando todas as possíveis novas dimensões, a partir da correlação entre dimensões originais. Com isso, é extraído o percentual de variância do sistema original capturada por cada uma das novas dimensões. Detalhes acerca da teoria dos métodos podem ser encontrados em Rodrigues (2021).

Usando a variância capturada por nova dimensão, determina-se quais são as mais significativas (distância até o cotovelo da distribuição). Selecionadas as dimensões mais relevantes, detalha-se qual a composição de cada uma em relação às milhares de dimensões originais. Esses milhares de componentes são agrupados por propriedade petrofísicas e é calculada a composição média de cada uma das propriedades.

Com as novas dimensões calculadas, o item (b) é executado, comparando as metodologias de clusterização Kmeans e Kmedoids em relação à quantificação de risco dos MRs selecionados. O MR é a realização geoestatística mais próxima ao centroide do cluster.

Como os métodos de clusterização mencionados não são determinísticos, cada seleção de MRs pode gerar um conjunto diferente. Assim, é necessário selecionar diversos conjuntos de MRs para entender o erro médio na quantificação de risco. Usando os seguintes números de novas dimensões: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 50, 100, 150, 200, 250, 350, 400, 450 e 500, a seleção de 12 MRs é realizada 10 vezes para cada um dos métodos de clusterização. Em cada uma das vezes, um novo conjunto de MRs é selecionado.

Considerando as realizações geoestatísticas e MRs equiprováveis, e usando uma estratégia de produção predefinida, são geradas curvas de risco para o indicador valor presente líquido (VPL). A curva de risco referência é gerada a partir da função de distribuição acumulada de todo conjunto de realizações geoestatísticas. Gera-se a curva de risco com os MRs usando regressão polinomial com condições de contorno, descrita por Rodrigues (2020).

A quantificação de risco é calculada a partir da comparação entre as duas curvas, usando o erro médio quadrático (RMSE, na sigla em inglês), descrito na seguinte equação:

ISSN: 2446-8339

$$RMSE = \sqrt{\sum_{i=1}^{m}\left(x_{ref,i} - x_{MRs,i}\right)^2 / m}$$

Valores médios de RMSE são calculados para cada metodologia de clusterização, considerando diferentes números de novas dimensões na seleção de MRs.

Para a execução do último item (c), a metodologia de clusterização para seleção de MRs é fixada em Kmeans, e varia-se o número de MRs selecionados em 3, 12 e 25. Desse modo, obtendo o RMSE médio para cada número de MRs, considerando diferentes números de novas dimensões na seleção dos conjuntos.

A metodologia descrita é aplicada ao modelo benchmark UNISIM-II-D, desenvolvido por Correia et al. (2015), que leva em consideração dados de produção do Pré-sal e dados sintéticos. Representa um campo de petróleo na fase de desenvolvimento, com apenas 516 dias de produção. Sua malha possui cerca de 65.000 blocos ativos com dimensões médias de 100 x 100 x 8m. Cada um possui valores relacionados a 5 propriedades: permeabilidade da matriz, permeabilidade da fratura, porosidade da matriz, porosidade da fratura e "*net-to-gross*". Assim, há aproximadamente 300.000 dimensões originais (combinações entre bloco e propriedade).

O caso contempla 500 realizações geoestatísticas (mapas de propriedades petrofísicas), representando possíveis distribuições espaciais das cinco propriedades, levando em consideração a distribuição de probabilidade de cada uma. Considera-se um conjunto hipotético de poços verticais produtores e injetores, distribuídos uniformemente, ilustrado na Figura 1.

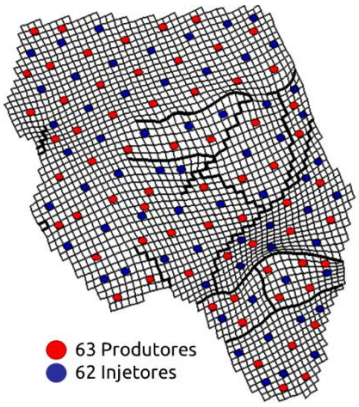

Figura 1 - Posicionamento dos poços produtores e injetores na estratégia utilizada.

RESULTADOS E DISCUSSÃO

ISSN: 2446-8339

Gerando o gráfico de variância capturada na forma de Pareto, verifica-se que a primeira dimensão gerada, ou também chamada de componente principal (PC, na sigla em inglês), é muito mais significativa que a última em termos de variância, sendo aproximadamente 40 vezes maior. Entretanto, ainda assim não é suficiente para capturar uma porção representativa da variabilidade do sistema, conforme apresentado na Figura 2.

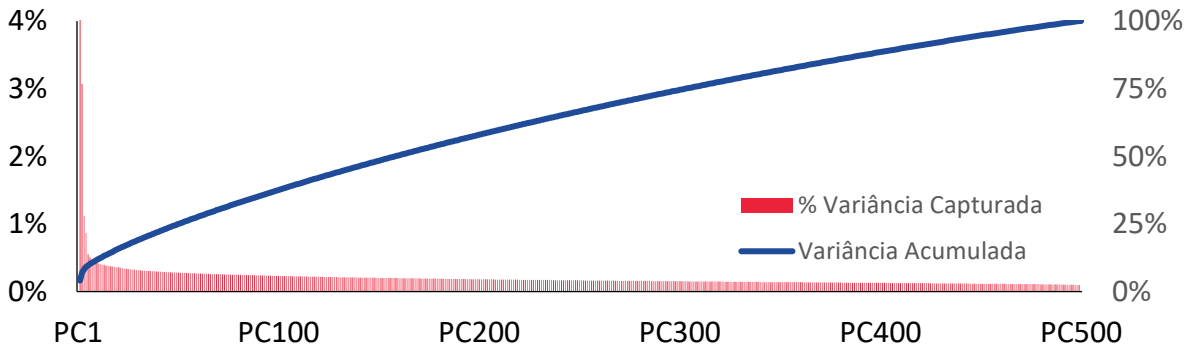

Figura 2 - Gráfico de Pareto da porcentagem de variância acumulada por PC.

A curva decai rapidamente, com a formação de um cotovelo na PC4. Após esse ponto, a variância por PC decai lentamente, indo de 0,3% na PC5 até 0,1% na PC500, o que é próximo a considerar que todas as PCs têm a mesma quantidade de informação ($1/500 = 0{,}2\%$). Analisando as primeiras PCs (Figura 3), nota-se que, até o cotovelo, a variância acumulada é de 9,1% e a redução de dimensões seria de 99,2% ($1 - 4PCs/500PCs$). Assim, apesar de capturar parte limitada das informações do sistema, é uma simplificação razoável.

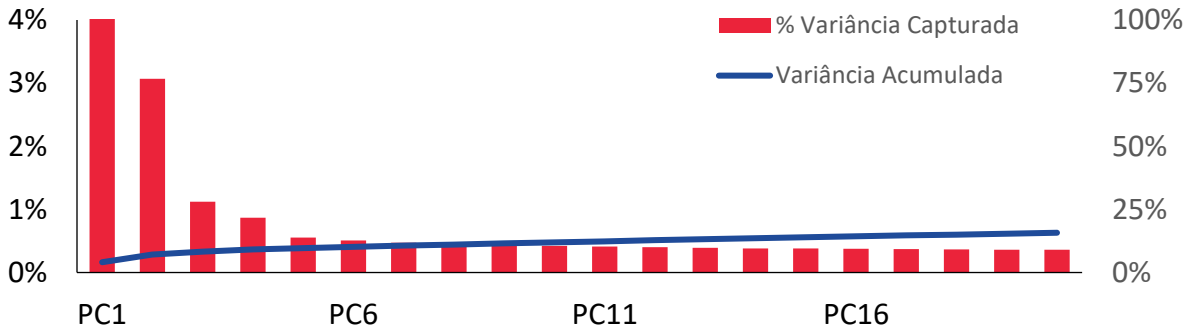

Figura 3 - Zoom no Gráfico de Pareto da porcentagem de variância acumulada por PC (PCs de 1 a 20).

A baixa presença de correlações é esperada para esse modelo de reservatório, visto que as propriedades das amostras são geradas através de sorteios que seguem a distribuição de um

ISSN: 2446-8339

variograma, e perfis de poços já perfurados. Assim, uma vez que UNISIM-II-D se encontra na fase de desenvolvimento, com poucos poços perfuradas e muita incerteza na distribuição espacial das propriedades petrofísicas, a dispersão do variograma é alta. Desse modo, a correlação entre as dimensões será baixa.

Avaliando a composição média das propriedades do reservatório nas 4 primeiras PCs, apresentada na Figura 4, observa-se que a PC de maior variância do conjunto de dados compreende majoritariamente dimensões relacionadas à porosidade da matriz. Desse modo, as maiores correlações do sistema acontecem entre as dimensões dessa mesma propriedade.

Na PC seguinte, nota-se que o valor de porosidade da matriz diminui. O que faz sentido, visto que pela teoria, todas as PCs são ortogonais, fazendo com que PC2 seja perpendicular à PC1. Os valores de fratura da permeabilidade e fratura da porosidade aumentam, indicando que, além de autocorrelação, existe correlação entre as dimensões dessas duas propriedades.

Seguindo, nota-se que em PC3 a composição é majoritariamente relacionada a matriz da permeabilidade, indicando que suas dimensões sejam mais correlacionadas entre si.

A propriedade "*net-to-gross*" não foi propriedade majoritária de nenhuma PC, não demonstrando relevância em suas autocorrelações ou correlações com outras propriedades.

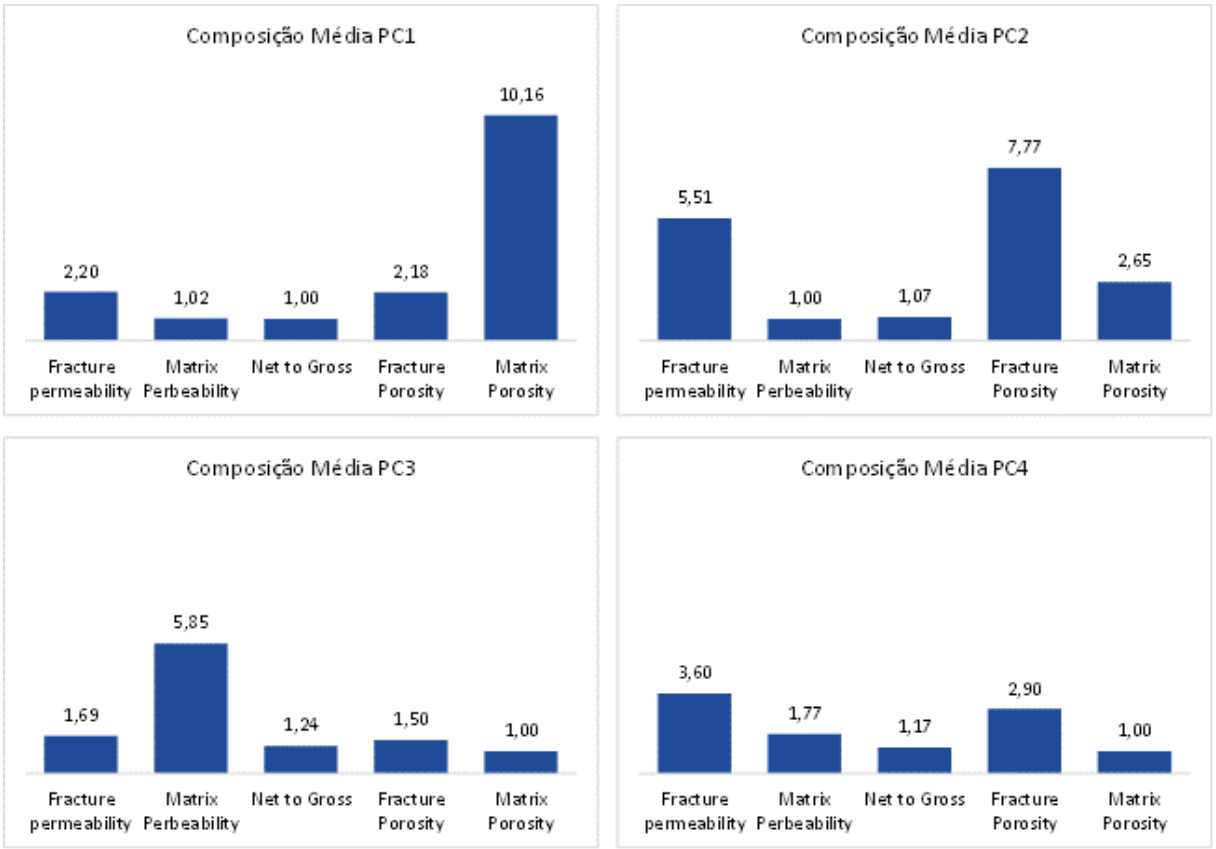

Figura 4 – Composição média das PCs mais significativas em termos de variância capturada.

 ISSN: 2446-8339

Comparando o espectro de "*net to gross*" com o de propriedades de alta correlação, como porosidade da matriz, verifica-se que ambas têm espectros muito distintos, apresentado na Figura 5. Enquanto porosidade da matriz possui um gradiente bem explorado de valores, "*net-to-gross*" é uma propriedade com valores binários, sendo 0, regiões de rocha não reservatório e 1, regiões de rocha reservatório. Segundo Correia et al. (2015) apenas 15% do mapa é do tipo não reservatório, tornando difícil criar correlações entre si e com as outras propriedades.

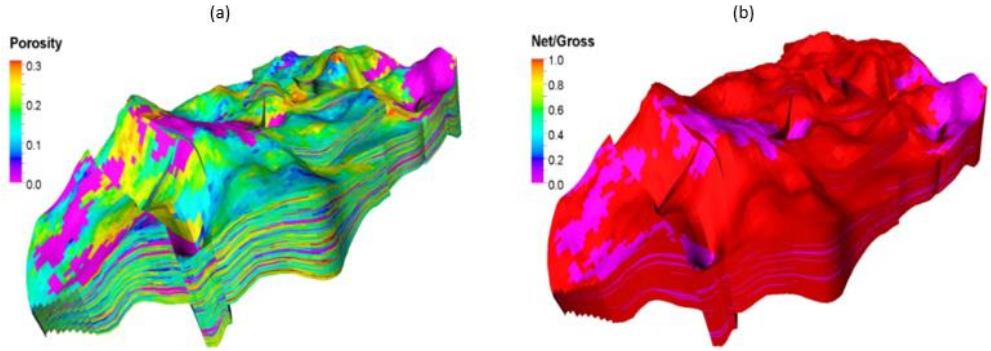

Figura 5 - Mapas de (a) porosidade da matriz e (b) "net-to-gross". Retirada de Correia et al. (2015).

Por último, ainda sobre a composição média das novas dimensões, a cada nova PC, o valor máximo da composição média diminui. Isso faz com que, a cada nova componente, as composições sejam cada vez mais equilibradas, fortalecendo a ideia de que após a PC4, as componentes passam a ser semelhantes, carregando a mesma quantidade de informação.

Os resultados das Figuras 2 a 4 revelam que poucas PCs não são suficientes para capturar a maior parte da variabilidade do conjunto de dados. Assim, cabe verificar se essa simplificação é válida para a que seleção de MRs gere boas estimativas na quantificação de risco.

Através do gráfico de RMSE médio pelo número de PCs usadas na clusterização (Figura 6), verifica-se que os erros associados às metodologias de Kmeans e Kmedoids são próximos e tendem a piorar conforme o número de PCs aumenta. Entretanto, Kmedoids oscila com uma amplitude maior, e conforme o número de PCs aumenta, verifica-se que Kmeans continua oscilando em torno de 10% de erro, enquanto Kmedoids oscila em torno de uma tendência crescente que sai dos mesmos 10%, mas chega aos 20%.

O comportamento associado ao aumento da RMSE à medida que o número de PCs aumenta segue o esperado pela "maldição da dimensionalidade", e é visto em Kmedoids. Contudo, não é claro motivo para que na curva de Kmeans, o efeito da "maldição" não seja pronunciado. Como em um espaço euclidiano, reduzir a distância de um ponto a uma direção é

 ISSN: 2446-8339

o mesmo que aumentar sua distância projetada a essa direção, PCA e Kmeans otimizam funções muito semelhantes. Assim, uma das hipóteses para esse comportamento, seria a similaridade entre essas duas metodologias.

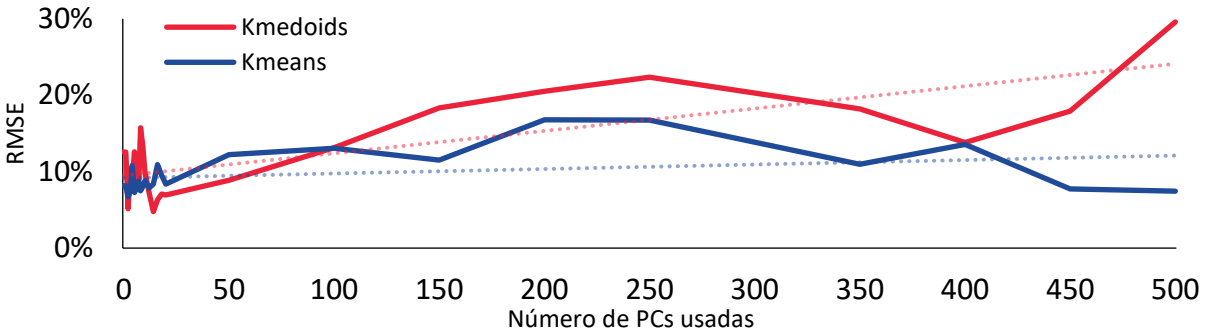

Figura 6 - RMSE médio por número de PCs usadas seleção de MRs a partir de Kmeans e Kmedoids.

Dessa maneira, Kmeans demonstra ser uma técnica melhor na representação de risco, já que, além de apresentar erros médio inferiores, é mais consistente em seus resultados.

Seguindo para a avaliação de como a quantificação de risco de conjuntos com diferentes números de MRs é afetada ao variar o número de PCs usadas na clusterização, obtêm-se o gráfico da Figura 7. As mesmas tendências da curva de Kmeans da Figura 6 se observam, mas verifica-se que, aumentando o número de MRs, o erro médio e amplitude de oscilação das curvas diminuem.

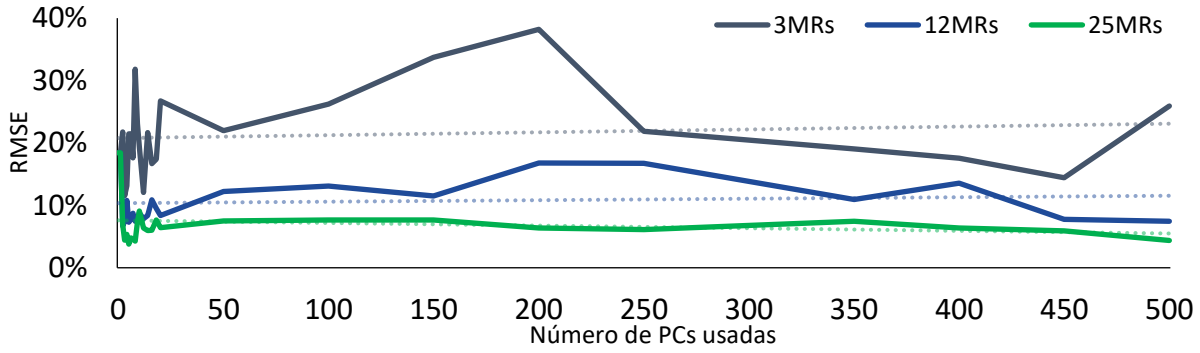

Figura 7 - RMSE médio por número de PCs usadas seleção de 3, 12 e 25MRs.

Nota-se que o número de MRs tem um grande impacto sobre a precisão da curva de risco, quantificada pelo RMSE, enquanto o número de PCs não parece ser muito relevante.

 ISSN: 2446-8339

Especialmente na curva de risco de 25 MRs, em que a curva do RMSE médio é próxima a uma reta horizontal.

Desse modo, por mais que seja interessante capturar o máximo de variabilidade com as PCs, é preferível simplificar o número de dimensões ao máximo para gastar mais processamento naquilo que é gargalo, o número de MRs na geração de curvas de risco.

CONSIDERAÇÕES FINAIS

No presente trabalho, a técnica de redução de dimensionalidades, PCA, foi usada para comprimir o conjunto de dimensões do modelo UNISIM-II-D, simplificando o sistema para seleção de MRs. Foram realizadas análises de sensibilidade quanto a quantidade de informação mantida após a redução de dimensionalidade e quantificação de risco, entendendo quanto os parâmetros e métodos de redução de dimensionalidade e clusterização afetam a seleção de MRs de realizações geoestatísticas. Foram variados o número de PCs, metodologia de clusterização e número de MRs para determinar a configuração que melhor representa risco.

Das análises referentes à quantidade de informação capturada pelo sistema reduzido, verifica-se que o modelo utilizado, UNISIM-II-D, não possui alta correlação entre suas dimensões. Com isso, a metodologia PCA não é capaz de concentrar a maior parte da informação do sistema em poucas dimensões. Seriam necessárias centenas de PCs para capturar mais da metade da variabilidade do sistema.

Considerando as PCs mais representativas, em termos de variância, seria possível capturar 9,1% da variabilidade em 4 PCs (redução de 99,2% do número de dimensões). O que é uma simplificação razoável, mas baixa, comparada a outras aplicações, como Wolf e Kirschner (2013) e Ngo (2018).

Em adição a isso, nota-se que a composição das PCs mais significativas é heterogênea em relação às propriedades que as dimensões representam. Dessa forma, caso o sistema simplificado use um baixo número de PCs, além de pouca informação, estará enviesado para as propriedades mais representativas na composição das PCs (porosidade da matriz, porosidade da fratura e permeabilidade da fratura).

Das análises referentes à quantificação de risco, verifica-se que a maioria das configurações usadas geram aproximações satisfatórias, com erros médios entre 6% e 20%. Contudo, diferente do estudo anterior, aumentar o número de PCs não melhora as aproximações, indicando que as informações mais relevantes para seleção de MRs estão nas

ISSN: 2446-8339

primeiras PCs, e que a redução de dimensionalidade favorece a clusterização, conforme a maldição da dimensionalidade explica.

Analisando a relação entre as técnicas de clusterização usadas e número de PCs, observa-se que o aumento do número de dimensões não favorece a quantificação de risco. Em ambos os casos da Figura 6, as curvas (que representam o erro médio da curva de risco gerada a partir de MRs) oscilam em torno de uma reta tendência. Entretanto, usando Kmeans, a curva oscila em torno de uma reta constante, enquanto usando Kmedoids, a curva oscila em torno de uma reta crescente. Dessa maneira, por conta do comportamento descrito, e por apresentar amplitude de oscilação inferior, Kmeans é uma metodologia de clusterização mais precisa para quantificação do risco.

Quanto à relação entre número de MRs e PCs, nota-se que configurações com diferentes números de MRs são afetadas da mesma forma pelo aumento do número de PCs, no que tange à quantificação de risco. Desse modo, tendo em vista evitar a maldição da dimensionalidade, continua sendo interessante reduzir o número de PCs.

Por último, o número de MRs se mostra como um parâmetro influente na representação de risco. Seu aumento beneficia a representação de risco, fazendo com que os MRs estejam mais bem distribuídos no comprimento da curva de risco, e o erro médio diminua. Nos experimentos realizados, observou-se a redução do erro médio de 21%, com 3MRs, a 6%, com 25MRs. Dessa forma, é interessante maximizar o número de MRs para obter estimativas mais precisas.

Assim, a partir das conclusões anteriores, a configuração do fluxo de trabalho que melhor representa risco é Kmeans, com 25 MRs e um baixo número de PCs (entre 2 e 4). Contudo, vale reforçar que quanto maior o número de MRs usado, melhor será a estimativa da curva de risco.

O trabalho foca na escolha de MRs e usa as curvas de risco para quantificar a representatividade dos modelos. Como sugestão de próximos passos é importante considerar diferentes aplicações de MRs. Além disso, dado que as metodologias de clusterização foram afetadas de formas diferentes pelo aumento do número de PCs, é interessante repetir as análises de quantificação de risco, usando outras metodologias de clusterização, tal como a DBSCAN, que é uma metodologia de clusterização determinística.

AGRADECIMENTOS

Agradecemos o apoio do EPIC - *Energy Production Innovation Center*, localizado na Universidade Estadual de Campinas (UNICAMP) e financiado pela Equinor Brasil Energia Ltda. e pela FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo (processo número 2017/15736-3). Agradecemos também o apoio da ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis) através do incentivo regulatório de P&D. Os agradecimentos estendem-se ao Centro de Estudos do Petróleo (CEPETRO) e à Faculdade de Engenharia Mecânica (FEM) da UNICAMP. Também reconhecemos o financiamento da Energi Simulation e agradecemos ao Computer Modeling Group Ltd. (CMG) pelas licenças e apoio de software.

REFERÊNCIAS

BELLMAN, Richard. *Dynamic Programming*. Princeton: Princeton University Press, 1957.

CORREIA, Manuel; Hohendorff, João; GASPAR, Ana Teresa; SCHIOZER, Denis. UNISIM-II-D: Benchmark case proposal based on a carbonate reservoir. In: SPE LATIN AMERICAN AND CARIBBEAN PETROLEUM ENGINEERING CONFERENCE, 2015, Quito, Ecuador. *Anais...* Richardson: Society of Petroleum Engineers, 2015. SPE-177140-MS. Disponível em: https://doi.org/10.2118/177140-MS. Acesso em: 19 abr. 2021.

MAHJOUR, Seyed Kourosh; SANTOS, Antonio Alberto Souza; CORREIA, Manuel Gomes; SCHIOZER, Denis José. Developing a workflow to select representative reservoir models combining distance-based clustering and data assimilation for decision making process. *Journal of Petroleum Science and Engineering*, v. 190, 107078, jul. 2020. Disponível em: https://doi.org/10.1016/j.petrol.2020.107078. Acesso em: 19 abr. 2021.

NGO, Linh. Principal component analysis explained simply. Bio Turing, [s. l.], 14 jun. 2018. Disponível em: https://blog.bioturing.com/2018/06/14/principal-component-analysis-explained-simply/. Acesso em: 2 jan. 2021.

RODRIGUES, Henrique Hungari. *Assessing the reliability of using representative models in risk analysis and decision-making processes during petroleum field development*. 2020. Relatório de Iniciação Científica (Graduação em Engenharia Mecânica) – Faculdade de Engenharia Mecânica, Universidade Estadual de Campinas, Campinas.

RODRIGUES, Henrique Hungari. *Avaliação do efeito da variação de dimensionalidade na seleção de realizações geoestatísticas representativas para quantificação de risco, através do método de Análise de Componente Principal*. 2021. 37 f. Trabalho de Conclusão de Curso (Graduação em Engenharia Mecânica) – Faculdade de Engenharia Mecânica, Universidade

 ISSN: 2446-8339

Estadual de Campinas, Campinas. Disponível em: https://www.unisim.cepetro.unicamp.br/publicacoes/HENRIQUE_HUNGARI_RODRIGUES.pdf. Acesso em: 19 abr. 2021.

SCHIOZER, Denis José; SANTOS, Antonio Alberto Souza; SANTOS, Susana Margarida Graça; HOHENDORFF FILHO, João Carlos von. Model-based decision analysis applied to petroleum field development and management. *Oil & Gas Science and Technology – Rev. IFP Energies nouvelles*, v. 74, 46, mai. 2019. Disponível em: https://doi.org/10.2516/ogst/2019019. Acesso em: 19 abr. 2021.

WOLF, Antje; KIRSCHNER, Karl N. Principal component and clustering analysis on molecular dynamics data of the ribosomal L11·23S subdomain. *Journal of Molecular Modeling*, v. 19, n. 2, p. 539-549, 2013. Disponível em: https://doi.org/10.1007/s00894-012-1563-4. Acesso em: 19 abr. 2021.