

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA DE PRODUÇÃO

METODOLOGIA DE PLANEJAMENTO E REDUÇÃO DO
NÚMERO DE EXPERIMENTOS EM PROBLEMAS DE BUSCA
ATIVA

Cláudia Eliane da Matta

Itajubá, 2021

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA DE PRODUÇÃO

METODOLOGIA DE PLANEJAMENTO E REDUÇÃO DO
NÚMERO DE EXPERIMENTOS EM PROBLEMAS DE BUSCA
ATIVA

Cláudia Eliane da Matta

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção como parte dos requisitos para obtenção do Título de Doutor em Ciências em Engenharia de Produção.

Banca Examinadora:

Prof. Dr. Claudimar Pereira da Veiga (Universidade Federal do Paraná – UFPR)

Prof. Dr. Wesley Vieira da Silva (Universidade Federal de Alagoas - UFAL)

Prof. Dr. Antonio Fernando Branco Costa (Universidade Federal de Itajubá - Unifei)

Prof. Dr. Pedro Paulo Balestrassi (Orientador)

Prof. Dra. Eliane Valença Nascimento De Lorenci (Orientadora)

Itajubá, 2021

DEDICATÓRIA

À Sofia Ruth da Matta Gomes Barbosa, minha amada filha, por me ensinar a ser uma pessoa melhor e por tornar minha caminhada cheia de significados.

AGRADECIMENTOS

Aos meus orientadores Pedro Paulo Balestrassi e Eliane Valença Nascimento De Lorenci, pela cuidadosa condução deste trabalho e pelos valiosos momentos de aprendizagem ao longo deste percurso acadêmico, pela felicidade de ter sido orientada por vocês e poder presenciar tantos exemplos de sabedoria, humildade, disposição e incentivos constantes.

Ao professor João Alves da Silva Neto, pela ajuda nos estudos sobre microrredes elétricas.

A minha família, Ricardo e Sofia, pela paciência durante todo esse longo percurso, pelo incentivo, amor e carinho constantes.

À minha mãe Lourdes, ao meu pai João (*in memoriam*), e às minhas irmãs Fabiana, Tânia e Edna, pelo apoio e carinho constantes no decorrer da minha vida.

Aos melhores amigos que alguém poderia ter: Denise, Jane, Juliana, Sandra e Eder por tantos momentos compartilhados.

Aos amigos que conheci durante o doutorado: Clarinha, Natália, Milena e Estevão, pela amizade e pela colaboração no percurso acadêmico.

Aos professores Zambroni, Bonatto, Paulo Ribeiro e Maurício, do grupo de pesquisa do qual faço parte, pelo acolhimento e aprendizagem constantes.

Ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Itajubá, pela estrutura oferecida ao Doutorado.

Por fim, tenho muito a agradecer a Deus por mais esta realização em minha vida.

EPÍGRAFE

Nunca ter tentado.

Nunca ter falhado.

Não importa.

Tentar outra vez.

Falhar outra vez.

Falhar melhor.

O fim está no início e ainda assim você segue em frente.

Samuel Beckett

RESUMO

Existem muitos problemas de engenharia que envolvem a otimização da função objetivo desconhecida. Recentemente, a busca ativa surgiu como uma ferramenta poderosa para resolver problemas dessa natureza, cujas funções objetivo envolvem alto custo de avaliação, seja este computacional ou experimental. Nesta proposta de doutorado busca-se encontrar um objeto (x) com valor ótimo para uma determinada propriedade (y). No entanto, a determinação direta desta propriedade de interesse em todos os objetos disponíveis pode não ser uma opção viável, tendo em vista os recursos, a carga de trabalho e/ou o tempo necessários. Dessa forma, este estudo propõe uma abordagem de aprendizado ativo de máquina, chamada busca ativa, destinado a encontrar uma solução ótima, utilizando delineamento de experimentos para busca inicial. Para aplicação do método foram utilizadas duas técnicas de regressão, chamadas de k -vizinhos-mais-próximos e processos Gaussianos. Além disso, um critério de parada foi definido para a técnica de regressão Gaussianiana, com o objetivo de reduzir o tempo de processamento do algoritmo. A originalidade do tema encontra-se na metodologia proposta, na utilização de delineamento de experimentos, no algoritmo de busca ativa usando técnicas de regressão que convergem rapidamente para um ótimo global e na utilização de um critério de parada para o algoritmo baseado em critérios estatísticos. Os estudos foram realizados com dados simulados e com dados reais para produção de medicamentos, agroquímicos e aplicação em microrredes elétricas. Em todos esses casos, a busca ativa reduziu o número de experimentos e simulações para obter a propriedade de interesse, em comparação com os algoritmos tradicionais, como o planejamento ótimo de experimentos e o Kennard-Stone.

Palavras-chave: Aprendizado de máquina. Busca ativa. K -vizinhos-mais-próximos. Processos Gaussianos. Planejamento de experimentos. Otimização.

ABSTRACT

Many engineering problems involve the optimization of the unknown objective function. Recently, active search has emerged as a powerful tool to solve problems of this nature, whose objective function involves high evaluation costs, whether computational or experimental. This thesis proposal seeks to find an object (x) with an optimal value for a given property (y). However, direct determination of this property of interest across all available objects may not be a viable option given the resources, workload and/or time required. Thus, this proposes an active machine learning approach, called active search, to find an optimal solution, using the design of experiments for the initial search. To apply this method, two regression techniques were used, called k -nearest-neighbours and Gaussian processes. Furthermore, a stopping criterion was defined for the Gaussian regression technique to reduce the algorithm processing time. The originality of the theme lies in the proposed methodology, in the use of experimental design, no active search algorithm using regression techniques that quickly converge to a global optimum, and in the use of a stopping criterion for the algorithm based on statistical criteria. The studies were carried out with simulated data and with real data for the production of medicines, agrochemicals and application in electrical microgrids. In all cases, active search reduced the number of experiments and simulations to obtain the property of interest, compared to traditional algorithms such as Optimal Experiment Design and Kennard-Stone.

Keywords: *Machine Learning. Active search. K-nearest neighbour. Gaussian processes. Design of experiments. Optimization*

LISTA DE FIGURAS

Figura 3.1. Visão geométrica do planejamento fatorial completo: (a) dois fatores e (b) três fatores	27
Figura 3.2. Experimentos fatoriais no Minitab®.....	30
Figura 4.1. Classificação da pesquisa.....	32
Figura 4.2. Bases de dados de periódico utilizadas para revisão sistemática.....	33
Figura 4.3. Relação entre os x -objetos e a propriedade y	34
Figura 4.4. Fluxograma para escolha do modelo de ajuste do ODoE.....	38
Figura 4.5. Objeto com a maior distância mínima do objeto já selecionado é indicado pela seta	40
Figura 4.6. Ajuste e predição do modelo GRP.....	45
Figura 4.3. Proposta de busca ativa para encontrar a solução ótima.....	47
Figura 4.8. Procedimentos metodológicos	48
Figura 5.1. Relação entre os x -objetos e a propriedade y no exemplo simulado	49
Figura 5.2. Resultado da busca com a utilização dos algoritmos ODoE e KS.....	50
Figura 5.3. Gráfico comparativo da busca ativa utilizando as técnicas k NNR e GPR com inicialização pelo ODoE.....	51
Figura 5.4. Resultados do método de busca ativa no primeiro exemplo simulado: (a) 6 objetos iniciais selecionados pelo algoritmo ODoE; (b) ponto ótimo encontrado no 15° objeto avaliado, utilizando k NNR; (c) ponto ótimo encontrado no 15° objeto avaliado, utilizando GPR; e (d) ponto ótimo encontrado no 86° objeto, utilizando GPR com critério de parada.	52
Figura 5.5. Gráfico comparativo da busca ativa utilizando as técnicas k NNR e GRP com o KS.	53
Figura 5.6. Resultados do método de busca ativa no primeiro exemplo simulado: (a) 6 objetos iniciais selecionados pelo algoritmo KS; (b) ponto ótimo após a seleção de 14 objetos utilizando k NNR; (c) ponto ótimo encontrado no 13° objeto avaliado, utilizando GPR; e (d) ponto ótimo encontrado no 13° objeto avaliado, utilizando GPR com critério de parada.	54
Figura 5.7. Gráfico comparativo da busca ativa com o ODoE.....	55
Figura 5.8. Gráfico comparativo da busca ativa com o KS.....	55
Figura 5.9. Resultados do método de busca ativa para o segundo exemplo simulado: (a) 6 objetos iniciais selecionados pelo algoritmo ODoE; (b) ponto ótimo encontrado no 21°	

objeto avaliado, utilizando k NNR; (c) ponto ótimo encontrado no 16° objeto avaliado, utilizando GPR sem critério de parada e (d) com critério de parada.....	56
Figura 5.10. Resultados do método de busca ativa no exemplo simulado: (a) 6 objetos iniciais selecionados pelo algoritmo KS; (b) ponto ótimo encontrado no 13° objeto avaliado, utilizando k NNR; (c) ponto ótimo encontrado no 14° objeto avaliado, utilizando GPR sem critério de parada; e (d) com critério de parada.....	57
Figura 5.11. Correlação entre as componentes do vetor \boldsymbol{x}_i na base QSAR Medicamento.	61
Figura 5.12. Resultado do método de busca ativa para a base de dados QSAR Medicamento, com a seleção dos objetos iniciais feito pelo KS.....	62
Figura 5.13. Resultado do método de busca ativa com a seleção dos objetos iniciais feito pelo ODoE.....	62
Figura 5.14. Ponto ótimo da base de dados QSAR Medicamentos utilizando seleção dos objetos iniciais pelo método KS e busca ativa com critério de parada para a técnica GPR.	63
Figura 6.15. Ponto ótimo da base de dados QSAR Medicamentos utilizando seleção dos objetos iniciais pelo método ODoE e busca ativa com critério de parada para a técnica GPR. ...	63
Figura 5.16. Correlação entre as componentes de \boldsymbol{X} de QSAR Toxicidade.....	66
Figura 5.17. Resultado do método de busca ativa para a base de dados QSAR Toxicidade, com a seleção dos objetos iniciais feito pelo KS.....	66
Figura 5.18. Resultado do método de busca ativa para a base de dados QSAR Toxicidade, com a seleção dos objetos iniciais feito pelo ODoE.....	67
Figura 5.19. Ponto ótimo da base de dados QSAR Toxicidade utilizando seleção dos objetos iniciais pelo método KS e busca ativa com critério de parada para a técnica GPR.	68
Figura 5.20. Ponto ótimo da base de dados QSAR Toxicidade utilizando seleção dos objetos iniciais pelo método ODoE e busca ativa com critério de parada para a técnica GPR. ...	68
Figura 5.21. Microrredes ilhadas do sistema principal.....	70
Figura 5.22. Correlação entre as componentes de \boldsymbol{X} que representam as microrredes.	75
Figura 5.23. Resultado do método de busca ativa considerando 32 possíveis agrupamentos de microrredes, com a seleção dos objetos iniciais feito pelo ODoE.	76
Figura 5.24. Resultado do método de busca ativa considerando 32 possíveis a agrupamentos de microrredes, com a seleção dos objetos iniciais feito pelo KS.....	77
Figura 5.25. Resultado do método de busca ativa considerando 31 possíveis agrupamentos de microrredes, com a seleção dos objetos iniciais feito pelo ODoE.	78
Figura 5.26. Resultado do método de busca ativa considerando 31 possíveis agrupamentos de microrredes, com a seleção dos objetos iniciais feito pelo KS.....	78

Figura 5.27. Agrupamento de microrredes após a aplicação do algoritmo de busca.	79
Figura 5.28. Redução do número de experimentos ou simulações com a implementação do método de busca ativa.....	81

LISTA DE QUADROS

Quadro 4.1. Algoritmo de Kennard-Stone	39
Quadro 4.2. Algoritmo de busca ativa utilizando a técnica k NNR	42
Quadro 4.3. Algoritmo de busca ativa utilizando a técnica GPR.....	46
Quadro 5.1. Descritores eletrônicos e suas definições	59
Quadro 5.2. Descritores eletrônicos e suas definições	65

LISTA DE TABELAS

Tabela 3.1. Matriz do planejamento fatorial completo para 3 fatores	28
Tabela 3.2. Efeitos principais e interações dado em função do número de fatores	29
Tabela 4.1. Bases de dados utilizadas para estudo de caso	35
Tabela 4.2. Número de componentes das bases de dados após análise da covariância.....	35
Tabela 5.1. Correlação entre as componentes de QSAR Medicamentos.	60
Tabela 5.2. Tempo médio de execução dos algoritmos para a base de dados QSAR Medicamento.	64
Tabela 5.3. Correlação entre as componentes do vetor x_i na base QSAR Toxicidade.....	66
Tabela 5.4. Tempo médio de execução dos algoritmos para a base de dados QSAR Toxicidade.	69
Tabela 5.5. Geração despachável e não-despachável das microrredes elétricas.	71
Tabela 5.6. Agrupamentos de microrredes no sistema proposto.	72
Tabela 5.7. Correlação entre as componentes definidas para representas as microrredes.	75
Tabela 5.8. Número de experimentos após a aplicação do algoritmo de busca	80
Tabela A.1. Estruturas moleculares e valores de pK_i para os 81 compostos arilpiperazínicos empregados nesta investigação.....	89

LISTA DE ABREVIATURAS E SIGLAS

ARD	<i>Automatic Relevance Determination</i>
BSA	<i>BusSearch Optimization Algorithm</i> - Algoritmo de otimização de busca
DoE	<i>Design of Experiments</i> - Planejamento de experimentos
GPR	<i>Gaussian Process Regression</i> - Regressão por processos Gaussianos
<i>k</i> NNR	<i>Nearest Neighbour Regression</i> - Regressão por <i>k</i> -vizinhos-mais-próximos
KS	<i>Kennard-Stone</i>
MARS	<i>Multivariate Adaptive Regression Spline</i>
ODoE	<i>Optimal Design of Experiments</i> - Planejamento ótimo de experimentos
RED	Recursos de Energia Distribuída
QSAR	<i>Quantitative Structure-Activity Relationships</i> - Relações quantitativas entre a estrutura química e a atividade biológica

SUMÁRIO

1	Introdução	16
1.1	Problema de pesquisa	17
1.2	Originalidade do tema e contribuições	17
1.3	Objetivos.....	18
1.4	Motivações sociais.....	18
1.5	Estrutura da tese.....	20
2	Fundamentação teórica	21
2.1	Notação preliminar	21
2.2	Problemas de otimização	21
2.3	Definição de busca ativa.....	23
2.4	Definição do problema de busca ativa.....	24
2.5	Trabalhos relacionados	24
3	Planejamento de experimentos.....	26
3.1	Arranjos fatoriais	26
3.2	Planejamento ótimo de experimentos.....	30
4	Procedimentos metodológicos.....	32
4.1	Caracterização da pesquisa.....	32
4.2	Exemplos simulados	33
4.3	Bases de dados reais	34
4.4	Método de busca ativa proposto	35
4.5	Algoritmo de busca inicial.....	36
4.6	Busca ativa.....	40
4.7	Proposta de busca ativa para encontra a solução ótima.....	47
5	Resultados e discussão da validação da metodologia	49
5.1	Primeiro exemplo simulado.....	49
5.2	Segundo exemplo simulado.....	54

5.3	QSAR Medicamento	58
5.4	QSAR Toxicidade	64
5.5	Aplicação em microrredes elétricas.....	69
5.6	Análises finais	79
6	Conclusões	83
	Referências	85
	Apêndice A – Estrutura Molecular do QSAR Medicamento	89

1 Introdução

O uso de simulações no campo da engenharia está cada vez mais recorrente, e cada vez mais necessário. As simulações visam a melhoria de processos e o desenvolvimento de novos produtos, com maior confiabilidade e qualidade. Geralmente processos experimentais, como por exemplo, a produção de uma vacina, o projeto aeroespacial, a síntese de proteínas *in vitro*, consomem muito tempo e recursos. Tais problemas buscam, via de regra, alguma forma de otimização e necessitam de simulação, envolvendo também um alto custo computacional (JIANG; MOSELEY; GARNETT, 2019).

Nessas aplicações, busca-se encontrar um objeto ou instância (aqui denominado por x), com valor ótimo para uma ou mais propriedades (denominada por y). Por exemplo, pode-se estar interessado em escolher um medicamento mais apropriado de uma certa família de compostos para obter o efeito terapêutico desejado, ou pode-se desejar reduzir o efeito de toxicidade de um agroquímico, ou ainda, pode-se desejar definir uma configuração ótima para uma rede de microrredes elétricas ilhadas com o objetivo de maximizar a capacidade de serviço. No entanto, nesses exemplos, a determinação direta de uma propriedade de interesse para todos os objetos disponíveis pode não ser uma opção viável, tendo em vista os recursos, a carga de trabalho e/ou o tempo necessários em um estudo experimental.

Sem uma base extensa de dados e sem a possibilidade de se fazer experimentos em abundância, isso representa um sério impedimento para a aplicação prática de algoritmos de otimização que estabelecem automaticamente os parâmetros críticos de projeto, em problemas do mundo real. Surge assim, o conceito de busca ativa, como uma alternativa neste cenário, no qual se deseja trabalhar com um pequeno conjunto de objetos em que suas propriedades (y) já sejam conhecidas e assim, estimar valores de y que estejam próximos ao valor ótimo para objetos cujas propriedades ainda não são conhecidas.

Quando as estimativas dessas propriedades podem ser obtidas de maneira indireta, usando parâmetros resultantes de cálculos teóricos ou medições instrumentais, pode-se elaborar modelos com base em um subconjunto de objetos de calibração com valores y conhecidos. A precisão das previsões do modelo tende a ser aprimorada à medida que mais objetos são adicionados ao conjunto de experimentos. No entanto, o uso de muitos objetos foge ao objetivo de usar métodos indiretos, que consistem em reduzir a carga de trabalho experimental, conseqüentemente do custo, associada à determinação direta da propriedade desejada. Idealmente, a busca pelos objetos com propriedades ótimas ou próximas ao ótimo deve ser realizada com o mínimo possível de determinações da propriedade de interesse.

1.1 Problema de pesquisa

O problema em encontrar um objeto ótimo ou próximo ao ótimo para uma função $y = f(\mathbf{x})$, utilizando um pequeno número de avaliações dessa função, tem atraído o interesse na literatura de otimização (GARNETT *et al.*, 2015; DAL BIANCO; GONCALVES; DUARTE, 2018; KIANI *et al.*, 2020). No entanto, esses trabalhos normalmente assumem que a função y pode ser avaliada para qualquer vetor \mathbf{x} , com quaisquer valores de componentes admissíveis. Em contraste, o presente trabalho refere-se a problemas nos quais as opções possíveis para \mathbf{x} são restritas a um conjunto finito de objetos disponíveis para seleção para se encontrar uma propriedade ótima. Geralmente as técnicas de inicialização de tais algoritmos de busca se baseiam em tentativa e erro, ou heurísticas aleatórias.

A partir deste contexto, tem-se a seguinte pergunta de pesquisa “Como é possível desenvolver uma abordagem estruturada para busca de valores ótimos utilizando inteligência artificial, baseada no paradigma de aprendizagem ativa de máquina?”.

1.2 Originalidade do tema e contribuições

Após a análise dos artigos que tratam sobre busca ativa (GHESU *et al.*, 2018; HÄSE; ROCH; ASPURU-GUZIŁ, 2019; SURAZHEVSKY; MINNEKHANOV; DEMIN, 2021), verificou-se que originalidade do tema está no método de busca proposto para encontrar um objeto ótimo em um conjunto finito de objetos disponíveis para seleção. Desta forma, a contribuição desta tese é apresentar um algoritmo de busca ativa usando técnicas de regressão que convergem rapidamente para esse ótimo global, se comparado a outras técnicas.

Primeiramente, é necessário definir um pequeno conjunto de objetos para identificar a propriedade de interesse por meio da experimentação. Para definição desse conjunto inicial é utilizada metodologia de delineamento de experimentos e, para efeito de comparação, é utilizado o algoritmo Kennard-Stone, um método clássico para seleção de objetos.

Após essa definição, é determinada a propriedade de interesse desse conjunto, em seguida, é aplicado o algoritmo de busca ativa. O algoritmo proposto inicia sua busca com esse pequeno conjunto de objetos com suas propriedades conhecidas (y). A cada iteração, o objeto com melhor valor estimado \hat{y} é então selecionado como provável candidato à otimização. O número de interações é definido pelo pesquisador e corresponde à n_{max} . Esse objeto candidato é então submetido ao procedimento de análise direta para avaliação da propriedade y . Após identificação da propriedade, esse fará parte do conjunto utilizado para seleção do próximo objeto.

São utilizadas duas técnicas de busca ativa: a regressão por k -vizinhos-mais-próximos e a regressão por processos Gaussianos. A primeira é uma técnica simples que geralmente apresenta um desempenho muito bom para uma ampla gama de problemas, especialmente quando o número de objetos é grande e há pouco ruído nos dados (LEON; CURTEANU, 2017) e a segunda técnica consiste em um modelo probabilístico, não paramétrico, para a função de regressão (BALLABIO *et al.*, 2019; SHAHRIARI *et al.*, 2016). A grande vantagem dessa última é que ela possui melhor desempenho preditivo quando comparada com outros métodos de regressão não paramétricos.

Para validação do método são apresentados dados simulados e estudos de caso reais. No primeiro estudo de caso real se está interessado em escolher o medicamento mais apropriado para o tratamento de depressão. O segundo visa encontrar a menor toxicidade de uma substância química. O último estudo envolve a determinação de um agrupamento ótimo para uma rede de microrredes ilhadas visando a maximização da capacidade de serviço.

A originalidade do tema encontra-se na metodologia proposta, na utilização de delineamento de experimentos, no algoritmo de busca ativa usando técnicas de regressão que convergem rapidamente para um ótimo global e na utilização de um critério de parada para o algoritmo baseado em critérios estatísticos.

1.3 Objetivos

O objetivo do trabalho é criar uma metodologia estruturada para busca de valores ótimos utilizando inteligência artificial, baseada no paradigma de aprendizagem ativa de máquina, com a utilização delineamento de experimentos.

Os objetivos específicos do trabalho são: (1) a aplicação de delineamento de experimentos e do algoritmo Kenard-Stone para seleção inicial dos objetos que apresentam as melhores propriedades de interesse, (2) desenvolvimento uma abordagem de busca ativa destinada a buscar objetos com propriedades ótimas, (3) implementar computacionalmente a técnica em um conjunto de dados simulados e em problemas reais, (4) implementar um critério de parada e (5) avaliar os benefícios da metodologia de busca ativa proposta.

1.4 Motivações sociais

Uma das motivações sociais deste trabalho é que a busca ativa pode ser usada para reduzir o estágio de testes em animais, aumentar a eficiência e evitar desperdícios. Muitos laboratórios usam testes com animais em pesquisas biomédicas (DOKE; DHAWALE, 2015; GUPTA;

SHARMA; KUMAR, 2018; NATH; DE; ROY, 2022; ONAWOLE et al., 2018; PETETTA; CICCOCIOPPO, 2020). O número de animais usados em pesquisas aumentou com o avanço da pesquisa e do desenvolvimento de medicamentos. Todos os anos, milhões de animais são usados em testes experimentais em todo o mundo. A dor, angústia e morte dos animais durante experimentos científicos têm sido debatidos há muito tempo. Além da grande preocupação com a ética, a experimentação animal tem outras desvantagens, como a demanda por mão de obra qualificada, os protocolos demorados e o alto custo (DOKE; DHAWALE, 2015).

O uso sistemático e difundido de animais para testes de toxicidade e avaliação de risco é um fenômeno relativamente recente. O primeiro teste de toxicidade realizado em nome de autoridades públicas nos Estados Unidos, na primeira década do século XX, não utilizou animais, mas voluntários humanos (ROWAN, 2015), no qual doze jovens do sexo masculino que foram sujeitos a experimentos de alimentação de 1902 a 1904 que envolviam benzoato, bórax e formaldeído. O uso de humanos deu lugar a um uso cada vez mais massivo de animais de laboratório, tanto para testes de segurança e avaliação de risco, quanto para pesquisas biomédicas (ROWAN, 2015). No entanto, há muitos sinais de que as abordagens dos testes de segurança estão mudando rapidamente (ROWAN, 2015).

O uso mundial de pesticidas químicos aumentou exponencialmente durante as últimas décadas, visto que essas aplicações são enormes em vários programas de proteção de cultivos e controle de vetores de doenças. Mais especificamente, os países em desenvolvimento e baseados na agricultura consomem quantidades muito maiores desses produtos químicos. Pesticidas químicos são as moléculas tóxicas projetadas para fins específicos, e suas práticas de aplicação não científicas podem causar riscos a espécies não-alvo, incluindo humanos e ecossistemas (BASANT; GUPTA, 2017). Devido às características persistentes, transfronteiriças e semivoláteis dos pesticidas sintéticos, uma fração significativa de sua dose aplicada pode ser detectada durante longos períodos como resíduos nas culturas e nos solos, bem como na atmosfera. A exposição de longo prazo a esses produtos químicos afeta adversamente os sistemas nervoso, endócrino, imunológico, reprodutivo, renal, cardiovascular e respiratório em humanos. Além disso, a exposição excessiva a esses produtos químicos é a razão para a extinção de várias espécies de aves e outras espécies animais durante as últimas décadas (BASANT; GUPTA, 2017). Em vista do exposto, várias agências regulatórias têm se esforçado para avaliar a toxicidade dos pesticidas químicos existentes e novos. Assim, para garantir a segurança dos produtos químicos, há uma necessidade de testes de toxicidade em espécies de teste aquáticas.

Alternativas à redução do uso de animais em pesquisas biomédicas e testes de segurança tem surgido. Uma delas é o uso de algoritmos inteligentes que reduzam o número de experimentos, como a busca ativa.

Outra motivação para esta tese é pensar em alternativas para a otimização do uso de recursos para a produção de energia elétrica. O aumento expressivo da produção de energia elétrica de forma descentralizada tem impactos diversos, de natureza elétrica, econômica, ambiental e social (MACMACKIN; MILLER; CARRIVEAU, 2021). Além disso, novas estruturas de energia emergentes, como redes inteligentes, reconfiguraram a estrutura de produção, distribuição e armazenamento (NOSRATABADI; HEMMATI; KHAJOUEI GHARAEI, 2021). Neste contexto, o planejamento e a operação dos sistemas de transmissão e distribuição de energia devem ser reformulados, por meio da aplicação de novas ferramentas e metodologias para otimizar esses recursos, de forma a acomodar a crescente inserção de recursos energéticos distribuídos na matriz elétrica (SPERSTAD; DEGEFA; KJØLLE, 2020). Nesta tese, a metodologia proposta de busca ativa visa reduzir a perdas de energia em agrupamentos de microrredes ilhadas.

1.5 Estrutura da tese

Com a finalidade de se conhecer o que está em desenvolvimento na área realizou-se uma pesquisa bibliográfica em diversas bases de dados; as considerações realizadas a partir dessa pesquisa são apresentadas no segundo capítulo. No terceiro capítulo é fornecido um referencial teórico sobre a técnica de planejamento de experimentos utilizada neste trabalho. Os procedimentos metodológicos e o método de busca ativa proposto que teve como fundamento o aprendizado ativo de máquinas é descrito no quarto capítulo. No quinto capítulo são expostos os resultados e é feita a discussão; são apresentados estudos com dados simulados, e em seguida são feitos estudos de caso com conjuntos de dados reais. Por fim, expõe-se as conclusões sobre a pesquisa e apresenta-se sugestões para trabalhos futuros.

2 Fundamentação teórica

Antes da fundamentação teórica, é necessário definir a notação preliminar que será utilizada no decorrer desta tese.

2.1 Notação preliminar

- Os vetores são denotados por letras itálicas em negrito, tais como \mathbf{x} .
- As matrizes são denotadas por letras maiúsculas em negrito, tais como \mathbf{X} .
- A norma Euclidiana do vetor \mathbf{x} é denotada por $\|\mathbf{x}\|$.
- Um conjunto é denotado por letra maiúscula, tal como N .
- Um conjunto vazio é representado pelo símbolo \emptyset .
- A união de dois conjuntos A e B é denotada por $A \cup B$.
- A diferença entre dois conjuntos A e B é denotada por $A \setminus B$, isto é, os elementos de $A \setminus B$ são aqueles que pertencem à A e não pertencem à B .
- A notação $\arg \max_{i \in \{1,2,\dots,N\}} f(i)$ é utilizada para indicar o argumento que maximiza $f(i)$.
- Um símbolo de chapéu ($\hat{}$) é usado para indicar um valor estimado.
- O \mathbf{x} -vetores associados aos objetos em questão são denotados por $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. Assume-se que diferentes objetos estão associados a diferentes \mathbf{x} -vetores, isto é, $\mathbf{x}_i \neq \mathbf{x}_j$ para $i \neq j$.
- As componentes de um \mathbf{x} -vetor (também chamadas de fatores ou variáveis) são denotadas por c_1, c_2, \dots, c_d .

2.2 Problemas de otimização

Para que seja possível a resolução de um problema de otimização real, inicialmente é necessária a criação de um modelo matemático que represente tal problema, sendo esta etapa chamada de modelagem. A modelagem cria um modelo, geralmente expresso como uma função matemática, chamada de função objetivo. A melhor solução da função objetivo pode ser expressa no sentido de máximo ou mínimo, dependendo do problema tratado. Vale ressaltar que as tarefas de maximização e minimização são trivialmente relacionadas, pois uma função $f(\mathbf{x})$ a ser maximizada pode ser representada como a minimização de $-f(\mathbf{x})$.

Neste sentido, a otimização pode ser definida como a ciência para determinar as melhores soluções para certos problemas que são formulados matematicamente que, em geral,

são modelos matemáticos da realidade. A otimização pode ser estudada como um ramo da matemática pura, mas tem aplicações em quase todos os ramos da ciência e tecnologia. Envolve o estudo de critérios de otimização para os mais diversos tipos de problema, a determinação de métodos e algoritmos de solução, o estudo da estrutura desses métodos, e experimentações computacionais com métodos de solução aplicadas tanto a problemas de teses quanto a problemas do mundo real (FLETCHER, 1987).

Uma extensa parte da literatura científica tem sido dedicada ao problema de otimizar uma função não linear $f(\mathbf{x})$ sobre um conjunto compacto A . No domínio da otimização, este problema é formulado conforme Equação 1:

$$\max_{\mathbf{x} \in A \subset \mathbb{R}^d} f(\mathbf{x}) \quad (1)$$

Normalmente, presume-se que a função objetivo $f(\mathbf{x})$ tenha uma representação matemática conhecida, seja convexa ou pelo menos tenha um baixo custo de avaliação. Apesar da influência da otimização clássica no aprendizado de máquina, muitos problemas de aprendizado não estão em conformidade com essas suposições (BROCHU; CORA; DE FREITAS, 2010).

A aprendizagem de máquina pode ser uma alternativa para otimização de funções que não tenham uma representação matemática conhecida, sejam elas convexas ou que apresentem alto custo de avaliação. Entretanto, muitos problemas de aprendizagem não estão de acordo com esses pressupostos e, muitas vezes, a avaliação de uma função objetivo é custosa ou mesmo impossível, e as propriedades de derivadas e convexidades são desconhecidas (BROCHU; CORA; DE FREITAS, 2010).

Uma das maneiras propostas na literatura para solução desse problema é utilizar a aprendizagem ativa de máquina que visa minimizar o número de avaliações de função objetivo e é aplicável em situações em que não existe uma expressão fechada para a função, mas se pode obter observações dessa função em valores amostrados (CHABANET; BRIL EL-HAOUZI; THOMAS, 2021; VAN HOUTUM; VLASEA, 2021). O objetivo do aprendizado ativo é reduzir o número de instâncias de treinamento a serem rotuladas pelo oráculo, ou seja, reduz a necessidade de consultar um especialista da área para rotular instâncias de um problema de classificação, diminuindo assim o custo empregado.

Quando a rotulagem dos dados de entrada é computacionalmente custosa, a aprendizagem ativa, uma área da aprendizagem de máquina, pode ser uma ferramenta muito

eficiente para os modelos de treinamento. Nesse caso, pode-se começar com uma quantidade menor de objetos para o treinamento do modelo, e um conjunto de testes no qual o algoritmo procura possíveis objetos a serem adicionados no conjunto de treinamento. Esses objetos possuem informações mais relevantes sobre o modelo e, desta forma, melhoram a previsão ou classificação. Assim, a aprendizagem ativa pode reduzir significativamente a quantidade de dados necessárias para o treinamento e, conseqüentemente, reduzir os esforços do processo de rotulagem e, desta forma, melhora a eficiência computacional (KIANI *et al.*, 2020; LI; DEL CASTILLO; RUNGER, 2020).

A aprendizagem ativa é baseada na hipótese de que um modelo treinado com um pequeno conjunto de objetos rotulados pode ter um desempenho tão bom quanto um treinado com um conjunto de dados onde todos os objetos são rotulados (DING *et al.*, 2020). O algoritmo escolhe um objeto mais informativo e consulta interativamente um especialista da área para rotular esse dado. Espera-se que esta informação adicional, escolhida ativamente, possa ser aproveitada para reduzir a incerteza associada aos rótulos de uma classe no conjunto de dados não rotulados. Com efeito, a rotulagem explícita de cada padrão é antecipada para fazer com que os rótulos de classe implícitos de alguns dos padrões não rotulados se tornem mais precisos (KOTHARI; JAIN, 2003).

Um método proposto na literatura é usar a aprendizagem ativa para encontrar os extremos das funções objetivo cuja avaliação é custosa (DAL BIANCO; GONCALVES; DUARTE, 2018). Essa proposta visa minimizar o número de avaliações da função objetivo e é aplicável em situações onde não existe uma expressão fechada para tal função, mas na qual as observações podem ser obtidas em valores amostrados. A aprendizagem ativa visa selecionar um conjunto reduzido e não redundante de pares para produzir um conjunto de treinamento com informações que representem fielmente um conjunto muito maior de dados não rotulados (DAL BIANCO; GONCALVES; DUARTE, 2018).

2.3 Definição de busca ativa

A busca ativa é uma particularização de aprendizagem ativa e visa encontrar objetos ótimos em um conjunto de possíveis candidatos (JIANG; MOSELEY; GARNETT, 2019).

Por exemplo, em muitos cenários reais, pode ser muito mais fácil coletar dados de entrada (objetos) do que obter suas saídas (y), as quais requerem um alto custo da ação humana. Por esta razão, a busca ativa visa construir modelos que exploram os dados de entrada na busca de valores ótimos com o intuito de reduzir os custos, o quanto possível. Muitos problemas reais

são desta forma, tais como: a detecção de fraudes, a descoberta de medicamentos e a recomendação de produtos, são alguns exemplos. Neste sentido, a identificação de um fraudador, a descoberta de um novo medicamento contra o câncer ou a venda de um produto pode ser medido em termos monetários (GARNETT *et al.*, 2012).

2.4 Definição do problema de busca ativa

Suponha que se tenha um conjunto finito de objetos $X \triangleq \{x_i\}$ e um subconjunto $X_{SEL} \subset X$, no qual cada objeto possui um valor y correspondente, chamados aqui de objetos rotulados, e um subconjunto $X_{\overline{SEL}} \subset X$ com objetos não rotulados, os quais não se conhece a propriedade de interesse y . Deseja-se selecionar ativamente, por meio do algoritmo de busca, uma sequência de objetos para maximizar uma determinada função.

2.5 Trabalhos relacionados

Um exemplo de aplicação da busca ativa é o desenho de medicamentos no qual procura-se identificar compostos que exibem atividade de ligação com um determinado alvo biológico entre milhões de candidatos, que não é uma tarefa trivial (GARNETT *et al.*, 2015).

Em outra proposta de busca ativa, com utilização da técnica de regressão por k -vizinhos-mais-próximos, foram efetuados estudos em duas bases de dados, na primeira visava-se escolher o medicamento mais apropriado de uma determinada família de compostos para obter um efeito terapêutico desejado e na segunda base visava-se a seleção de espécimes de plantas com características fenotípicas adequadas para programas de reprodução (MATTA *et al.*, 2016).

Outras aplicações incluem recomendação de produtos (VANCHINATHAN *et al.*, 2015) e descoberta de materiais (JIANG *et al.*, 2018).

No setor elétrico, um método de otimização chamado algoritmo de otimização de busca (*Search Optimization Algorithm - BSA*) foi proposto para resolver o problema de fluxo de potência ótimo (CHAIB *et al.*, 2016). Este método foi testado para 16 casos diferentes nos sistemas de teste IEEE 30-bus, IEEE 57-bus e IEEE 118-bus. Os resultados obtidos foram comparados com aqueles obtidos usando alguns algoritmos de otimização conhecidos. Tal estudo destaca a eficácia do método para resolver diferentes problemas de fluxo de potência ótimo com funções objetivo complexas.

A busca ativa com uma abordagem de cadeia de Markov foi utilizada para o desenho de drogas, com o objetivo de encontrar medicamentos novos e eficazes para o tratamento de uma doença pulmonar crônica (OGLIC *et al.*, 2018). De acordo com esses autores, nos últimos 15

anos, a descoberta de medicamentos por meio de aplicações que utilizam o aprendizado de máquina tem se popularizado.

A detecção rápida e robusta de estruturas anatômicas de um paciente representa um componente importante das tecnologias de análise de imagens médicas. Uma solução para detecção de marcos anatômicos de dados volumétricos da região corporal de pacientes é proposta utilizando o princípio busca ativa. A velocidade de detecção do método é até 30 vezes mais rápida do que a referência de última geração de algoritmos, atingindo desempenho em tempo real (GHESU *et al.*, 2018).

Um estudo feito dentro do paradigma de aprendizagem ativa, chamado de busca ativa econômica, com o objetivo de encontrar um determinado número de pontos positivos em um grande conjunto de objetos sem rótulo, com custo mínimo de rotulagem, foi proposto utilizando uma abordagem Bayesiana (JIANG; MOSELEY; GARNETT, 2019).

Uma proposta de laboratórios autônomos nos quais métodos de inteligência artificial buscam ativamente procedimentos experimentais promissores, formulando hipóteses sobre seus resultados com base em experimentos anteriores é proposta por (HÄSE; ROCH; ASPURU-GUZIK, 2019).

Outra proposta de busca ativa para melhorar a compreensão da formação de defeitos em peças construídas com manufatura aditiva e controlar a variabilidade do processo em tempo real é proposta por (CHEN; IMANI; IMANI, 2021).

Esses trabalhos auxiliaram na elaboração da metodologia proposta nesta tese. Nesse sentido, o primeiro passo é a seleção dos objetos iniciais que serão utilizados na aquisição da propriedade de interesse (y). Uma das formas de realizar essa seleção é por meio do planejamento de experimentos, detalhado no próximo capítulo.

3 Planejamento de experimentos

O planejamento de experimentos (*Design of Experiment* - DoE) é uma abordagem poderosa utilizada para melhorar um processo (MONTGOMERY, 2013). É usada para encontrar uma solução ótima de um processo, descobrir soluções robustas (ASFAW *et al.*, 2020; KIRKEY *et al.*, 2020; YONDO; ANDRÉS; VALERO, 2018) e pode também ser aplicado a problemas de simulação para reduzir o trabalho computacional. Muitos problemas de engenharia requerem muitas simulações para desenvolver uma solução apropriada, o que exige um alto custo computacional (ALIZADEH; ALLEN; MISTREE, 2020).

Essa técnica envolve uma série de testes no quais um conjunto de componentes de entrada, também chamados de fatores, são alterados pelo experimentador de forma controlada para observar e identificar as razões para as mudanças na resposta de saída.

Desenvolvido em 1926 por Ronald A. Fisher, o DoE foi utilizado pela primeira vez para organizar experimentos agrícolas de campo (FISHER, 1992). O DoE é baseado em princípios estatísticos semelhantes aos da análise de variância (ANOVA) e análise de regressão (YONDO; ANDRÉS; VALERO, 2018). No planejamento de qualquer experimento, o primeiro passo é decidir quais são os fatores para se obter as respostas de interesse (y).

A definição do número de fatores (termo comumente utilizado por estatísticos), aqui chamados de componentes do objeto, é importante, pois um grande número de componentes vai consumir muito mais tempo e ter um maior custo operacional. Essa definição se dá por meio do arranjo fatorial completo ou fracionado.

Há diferentes abordagens sobre planejamento de experimentos, nas próximas seções serão definidas algumas dessas abordagens.

3.1 Arranjos fatoriais

Os arranjos fatoriais (MONTGOMERY, 2013) são um conjunto de combinações de níveis que permitem ao experimentador estudar o efeito conjunto das variáveis (componentes) de projeto nas funções objetivo. Em experimentos fatoriais, todas as combinações possíveis dos níveis das variáveis de projeto são investigadas durante cada replicação do experimento.

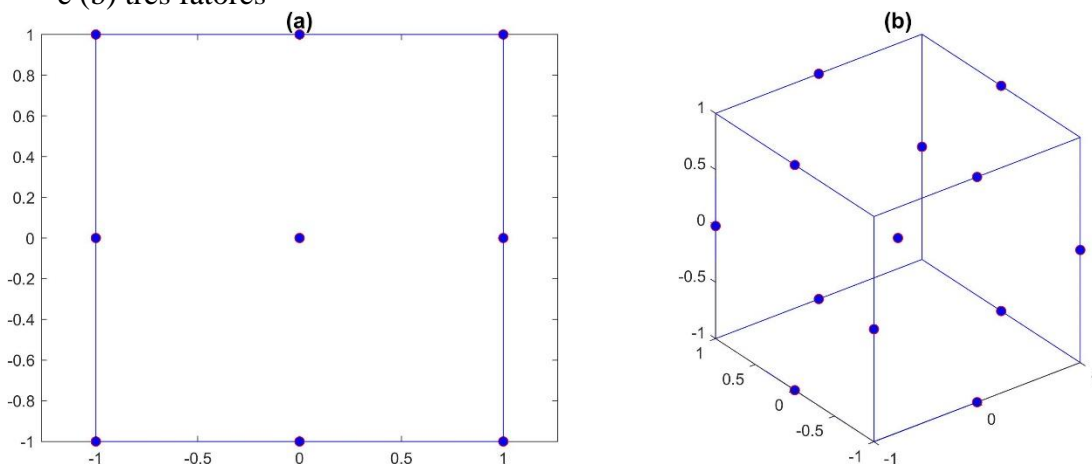
3.1.1 Arranjos fatoriais completos

Para fazer um planejamento fatorial completo, deve-se realizar experimentos em todas as possíveis combinações dos níveis dos fatores. Cada um desses experimentos é um ensaio experimental.

Um planejamento completo de dois níveis exige a realização de 2^k ensaios diferentes, sendo chamados por isso de planejamento fatorial 2^k , pois para k fatores, temos k variáveis controladas pelo experimentador.

Em um arranjo fatorial completo todos os fatores são utilizados para se obter as respostas de interesse. As Figuras 3.1 (a) e (b) apresentam a representação gráfica do experimento para exemplos de arranjos fatoriais 2^k completos para dois e três fatores, respectivamente. O planejamento experimental é representado em um sistema cartesiano, com um eixo para cada fator, no qual os vértices dos quadrados são os ensaios, os efeitos principais são contrastes (diferenças médias) entre valores situados em arestas opostas e perpendiculares ao eixo do fator correspondente e o efeito de interação é o contraste entre as duas diagonais, considerando-se positiva a diagonal que liga o ensaio (-1 -1) ao ensaio (1 1) ou (-1 -1 -1) ao ensaio (1 1 1), para dois e três fatores, respectivamente. Os valores dos efeitos foram todos escalados para ir entre -1 e 1, tornando os eixos adimensionais. A Figura 3.1 (a) apresenta dois fatores, o espaço definindo é um plano e a Figura 3.1 (b) apresenta três fatores, na qual o espaço é definido como um cubo.

Figura 3.1. Visão geométrica do planejamento fatorial completo: (a) dois fatores e (b) três fatores



Fonte: A autora.

As possíveis combinações podem ser apresentadas em uma tabela, chamada matriz de planejamento. Como exemplo, a Tabela 3.1 apresenta a matriz de planejamento de um arranjo

fatorial 2^k completo para três fatores. A nomenclatura indica que o fator aparece no ponto experimental com nível alto (+1).

Tabela 3.1. Matriz do planejamento fatorial completo para 3 fatores

Ensaio	A	B	C	Labels	A	B	C
1	-1	-1	-1	(1)	0	0	0
2	+1	-1	-1	a	1	0	0
3	-1	+1	-1	b	0	1	0
4	+1	+1	-1	ab	1	1	0
5	-1	-1	+1	c	0	0	1
6	+1	-1	+1	ac	1	0	1
7	-1	+1	+1	bc	0	1	1
8	+1	+1	+1	abc	1	1	1

Fonte: A autora.

O arranjo fatorial completo torna-se inviável, conforme o número de fatores e níveis aumentam, e intuitivamente, muitas vezes é desnecessário cobrir completamente os espaços de estados (região de busca) para obter a informação útil (ou propriedade de interesse). Conforme o número de fatores aumenta, a redundância do planejamento também aumenta (LEE, 2019). Desta forma, recomenda-se utilizar esta abordagem quando se têm até quatro fatores (YONDO; ANDRÉS; VALERO, 2018). Quando existem muitos fatores, utilizar o fatorial completo se torna inviável, pois aumenta muito o número de experimentos. Neste sentido, quanto maior o número de dimensões (variáveis de projeto), mais difícil se torna o problema experimentalmente. Assim, para cinco ou mais fatores pode-se utilizar planejamento fatorial fracionado que irá reduzir significativamente o número de experimentos.

O princípio da hierarquia afirma que os efeitos de ordem inferior tendem a dominar, o que significa que, na prática, os efeitos de terceira ou ordem superior podem ser considerados insignificantes, pois os esforços experimentais são altos e, provavelmente, a solução do problema poderia ter sido tratada com menos execuções experimentais (LEE, 2019).

Comumente, usa-se a triagem para identificar os fatores mais importantes para otimização, quando o número de fatores é igual ou superior a cinco. Em tais casos, os arranjos fatoriais completos são invariavelmente caros demais. A triagem é sinônimo da análise estatística com modelagem dos efeitos de primeira ordem, resultando em modelos de efeitos principais, embora alguns projetos de triagem permitam uma estimativa limitada de outros efeitos (LEE, 2019).

Existem vários tipos de planejamento para triagem de fatores, tais como: arranjos fatoriais fracionado, que incluem apenas um subconjunto do planejamento fatorial completo;

planejamento de Plackett-Burman; delineamento de compósito central e delineamento Box-Behnken.

3.1.2 Arranjos fatoriais fracionados

Os arranjos fatoriais fracionados consistem em subconjuntos ou frações de arranjos fatoriais completos (FINNEY, 1945). Com essa técnica, é possível analisar os efeitos sobre uma resposta de interesse, de 2^k combinações (em que k representa o número de componentes do objeto) para 2^{k-p} combinações, na qual p representa o nível de fatores (no exemplo apresentado na Tabela 3.1, da subseção anterior) são dois níveis, então $p = 2$), utilizadas para realizar a experimentação. Ou seja, com essa técnica, realiza-se apenas parte do experimento, sem comprometer significativamente a precisão das conclusões decorrentes da análise de resultados. Simultaneamente, os custos e o tempo de duração dos testes são significativamente reduzidos.

Sabe-se que com uma expansão em série de uma função, os efeitos principais (de primeira ordem) tendem a ser maiores do que as interações de dois fatores (segunda ordem), que por sua vez são mais importantes que as interações de ordem mais alta. Se esses efeitos não são significativos, determinar seu valor não é um motivo para se realizar todos os ensaios do planejamento completo (BARROS NETO; SCARMINIO; BRUNS, 2010). O número de efeitos principais e de interações (dado em função do número de fatores) são mostrados na Tabela 3.2.

A ordem de uma interação é o número de fatores envolvidos na sua definição. Por exemplo, se o objeto a ser analisado possui 3 componentes, os efeitos de primeira ordem são 3 (a, b, c), os efeitos de segunda ordem também são 3 (ab, ac, bc) e há um efeito de 3ª ordem (abc).

Tabela 3.2. Efeitos principais e interações dado em função do número de fatores

k (fatores)	2^k	Ordem						
		1 ^a	2 ^a	3 ^a	4 ^a	5 ^a	6 ^a	7 ^a
3	8	3	3	1	-	-	-	-
4	16	4	6	4	1	-	-	-
5	32	5	10	10	5	1	-	-
6	64	6	15	20	15	6	1	-
7	128	7	21	35	35	21	7	1

Fonte: Adaptado de (BARROS NETO; SCARMINIO; BRUNS, 2010).

Desta forma, presumem-se que certas interações de ordem superior não são significativas, então, uma avaliação razoável da influência das variáveis de projeto na função

objetivo e os efeitos de interação desejados pode ser obtida executando apenas um número mínimo de casos experimentais, portanto, uma fração do experimento de fato completo.

Quando se tem poucos fatores (até quatro), utiliza-se todas as componentes para criação do modelo, caso contrário, pode-se utilizar o arranjo fatorial fracionado para a escolha dos fatores influentes para construção do modelo. Nesta tese, a escolha do número de fatores foi feita por meio do Minitab®, em uma ferramenta específica para DoE, conforme Figura 3.2.

Figura 3.2. Experimentos fatoriais no Minitab®.

Criação de um Experimento Fatorial: Exibir Experimentos Disponíveis

Experimentos Fatoriais Disponíveis (com Resolução)

Ensa	Fatores													
	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4	Com	III												
8		Com	IV	III	III	III								
16			Com	V	IV	IV	IV	III	III	III	III	III	III	III
32				Com	VI	IV	IV	IV	IV	IV	IV	IV	IV	IV
64					Com	VII	V	IV	IV	IV	IV	IV	IV	IV
128						Com	VIII	VI	V	V	IV	IV	IV	IV

Resolução Disponível III Experimentos Plackett-Burman

Fatores	Ensaio	Fatores	Ensaio	Fatores	Ensaio
2-7	12,20,24,28,...,48	20-23	24,28,32,36,...,48	36-39	40,44,48
8-11	12,20,24,28,...,48	24-27	28,32,36,40,44,48	40-43	44,48
12-15	20,24,28,36,...,48	28-31	32,36,40,44,48	44-47	48
16-19	20,24,28,32,...,48	32-35	36,40,44,48		

Fonte: Minitab®.

Ao criar um experimento fatorial fracionado, o Minitab® utiliza a fração principal por padrão. A fração principal é a fração em que todos os sinais possuem nível alto (+1). As células destacadas em verde significam que este é um experimento viável, por exemplo, se o objeto tiver 5 componentes, pode-se iniciar a experimentação com 8 objetos ($2^{k-p} = 2^{5-2}$). Se fosse considerado o fatorial completo, seriam 32 objetos (2^5).

3.2 Planejamento ótimo de experimentos

O planejamento ótimo de experimentos (ODOE) é um método geral e flexível para aplicações nas quais o planejamento de experimentos clássico não se aplica. Ao contrário do planejamento de experimentos usualmente utilizado, como fatoriais completos ou fracionados, as matrizes de um planejamento ótimo de experimentos geralmente não são ortogonais e as estimativas de efeito são correlacionadas (MONTGOMERY, 2013).

O objetivo do ODoE é desenvolver planejamentos experimentais que sejam otimizados a partir de critérios estatísticos relacionados às estimativas dos parâmetros do modelo ou às previsões do modelo (LI; DEL CASTILLO; RUNGER, 2020). De acordo com esses autores, existem muitas analogias e semelhanças entre a aprendizagem ativa e os métodos sequenciais para ODoE.

O planejamento de experimentos de uma região cúbica ou esférica é solucionado por uma abordagem de planejamento de superfície. No entanto, pode-se encontrar uma situação na qual o planejamento de superfície de resposta padrão pode não ser a escolha óbvia. Nesse caso, o ODoE é uma alternativa a ser considerada (MONTGOMERY, 2013).

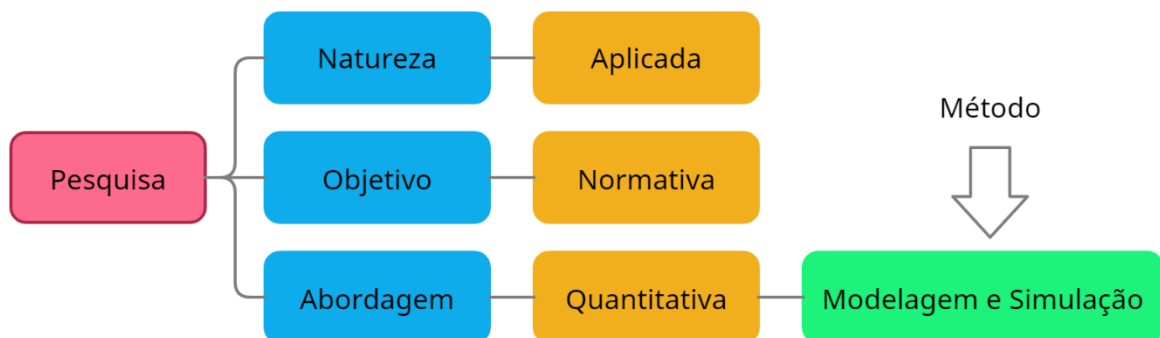
4 Procedimentos metodológicos

Os procedimentos metodológicos, o método de busca proposto, os exemplos simulados e as bases de dados reais são apresentados nas próximas seções.

4.1 Caracterização da pesquisa

Esta é uma pesquisa de natureza aplicada que se caracteriza por seu interesse prático, pois os resultados podem ser aplicados ou utilizados imediatamente na solução de problemas reais. É uma pesquisa normativa com o objetivo de aperfeiçoar os resultados disponíveis na literatura existente. Sua abordagem é quantitativa. A modelagem e simulação são empregadas, pois se deseja realizar experimentações de um sistema real por meio de um modelo. A Figura 4.1 mostra a classificação desta pesquisa.

Figura 4.1. Classificação da pesquisa

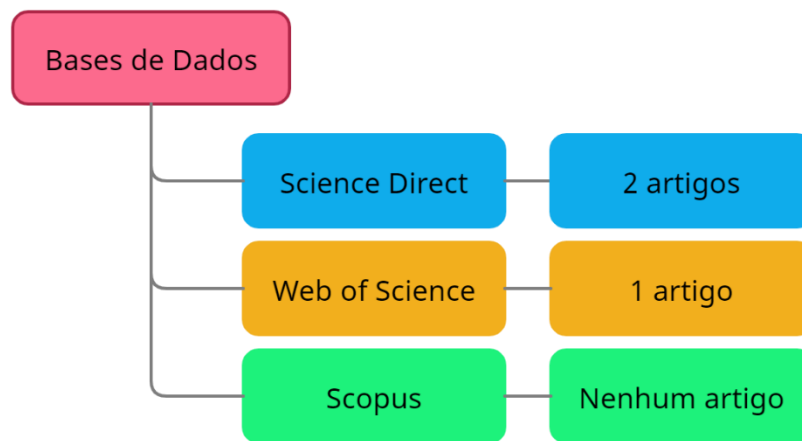


Fonte: A autora.

Uma busca sistemática foi realizada em periódicos indexados nas bases de dados que estão dentro da área de conhecimento de Engenharia de Produção, de acordo com o portal de Periódicos da Capes. Dentre as bases apresentadas, foram selecionadas a *Web of Science*, a *ScienceDirect* e a *Scopus*. Os critérios de inclusão dos artigos foram os seguintes: conter descritores escolhidos: “*active search*” AND “*machine learning*”, publicados em inglês, entre 2016 e 2021. Os critérios de exclusão foram resumos de conferências ou capítulos de livros. Todos os resumos dos artigos encontrados foram analisados para determinar ou não a inclusão na tese. Essa delimitação do escopo visa obter artigos recentemente publicados sobre o assunto “busca ativa”.

Na base de dados “*ScienceDirect*” foram encontrados dois artigos sobre o assunto (GHESU et al., 2018; HÄSE; ROCH; ASPURU-GUZIŁ, 2019), na “*Web of Science*” foi encontrado um artigo (CHEN; IMANI; IMANI, 2021) e na Scopus nenhum artigo foi encontrado. Esses artigos foram detalhados no capítulo sobre a fundamentação teórica. A Figura 4.2 ilustra a revisão sistemática realizada.

Figura 4.2. Bases de dados de periódico utilizadas para revisão sistemática.



Fonte: A autora.

Os algoritmos apresentados neste trabalho foram implementados no *software* MatLab® R2020a. Para definição de n_o foi utilizado a ferramenta para o planejamento de experimentos, disponível no *software* Minitab® 2018.

Nas próximas seções estão descritas as bases de dados utilizadas nesta tese.

4.2 Exemplos simulados

O estudo inicial utilizou dois exemplos simulados para verificar a efetividade do método de busca proposto.

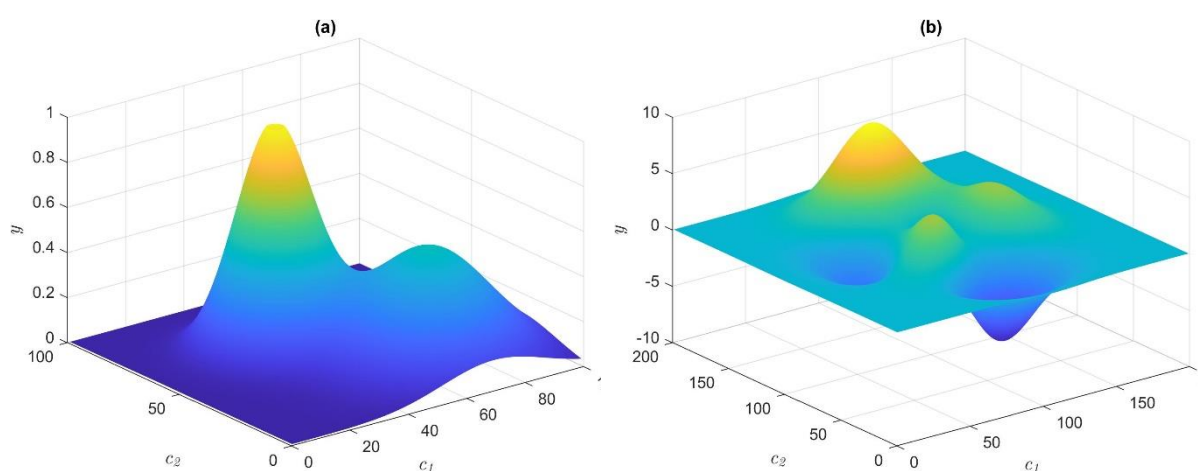
O primeiro exemplo, envolveu duas componentes (c_1 e c_2) relacionadas a uma propriedade y de acordo com a Equação 9. Essa expressão gera uma superfície com máximos locais e globais, como pode-se observar na Figura 4.3 (a).

$$y = e^{-\left(\frac{c_1-40}{15}\right)^2 - \left(\frac{c_2-60}{15}\right)^2} + \frac{1}{2} e^{-\left(\frac{c_1-70}{25}\right)^2 - \left(\frac{c_2-30}{25}\right)^2}, \quad (2)$$

O segundo exemplo simulado, similar ao primeiro, também envolveu duas componentes (c_1 e c_2) relacionadas a uma propriedade y , de acordo com a Equação 10. No entanto, essa expressão gera uma superfície com máximos e mínimos globais e locais, diferentemente do primeiro exemplo, conforme Figura 4.3 (b).

$$y = 3(1 - c_1)^2 e^{-c_1^2 - (c_1+1)^2} - 10 \left(\frac{c_1}{5} - c_1^3 - c_2^5 \right) e^{-c_1^2 - c_2^2} - \frac{1}{3} e^{-(c_1+1)^2 - c_2^2}, \quad (3)$$

Figura 4.3. Relação entre os x -objetos e a propriedade y .



Fonte: A autora.

No primeiro exemplo simulado um total de 150 objetos são gerados para compor pares randômicos (c_1, c_2), com valores entre $[0,100]$ e no segundo exemplo são gerados 200 objetos. O problema consiste em achar o objeto mais próximo do máximo global de y .

4.3 Bases de dados reais

Nesta proposta de tese, são estudados duas bases de dados reais. A primeira base refere-se às relações quantitativas entre a estrutura química e a atividade biológica (QSAR) para desenvolvimento de um medicamento. A segunda base de dados refere-se às QSAR de moléculas orgânicas que preveem a toxicidade aquática aguda. Além disso, é simulada uma aplicação em microrredes elétricas ilhadas, em que cada microrrede é derivada de adaptações do sistema de distribuição teste IEEE-37 barras.

O número total de objetos, o número de amostra de cada subconjunto, o número de componentes e o número de experimentos iniciais são apresentadas na Tabela 4.1.

Tabela 4.1. Bases de dados utilizadas para estudo de caso

Base de dados	Número de objetos		Número de Componentes	Experimentos iniciais
	Total	Tamanho da amostra		
QSAR Medicamento	079	050	14	16
QSAR Toxicidade	908	100	06	08
Microrredes	032	032	12	08

Fonte: A autora.

Foi feita a análise de covariância das bases de dados. Componentes dos objetos com alta correção (p -value > 0,005) foram retiradas. Após esta análise, o número de componentes das bases de dados de QSAR Medicamentos reduziu para 10 e do agrupamento de Microrredes para 5, conforme Tabela 4.2.

Tabela 4.2. Número de componentes das bases de dados após análise da covariância.

Base de dados	Número de Componentes
QSAR Medicamento	10
QSAR Toxicidade	06
Microrredes	05

Fonte: A autora.

Foram gerados 100 subconjuntos diferentes, utilizando um procedimento de subamostragem, com 50 amostras para QSAR medicamento e 100 amostras para o QSAR toxicidade, a partir do número total de objetos de base de dados gerados por meio de distribuição uniforme. O método de busca ativa proposto foi então aplicado a cada um destes subconjuntos com o objetivo encontrar o objeto com o maior valor de y em cada subconjunto.

Para aplicação em microrredes elétricas foi proposta uma metodologia para definição dos objetos e não foram gerados subconjuntos.

4.4 Método de busca ativa proposto

O método de busca ativa proposto teve como fundamento o aprendizado ativo, uma área emergente no aprendizado de máquina que participa ativamente para a escolha de exemplos de treinamento, permitindo assim maximizar a taxa de acerto para uma dada quantidade de objetos rotulados (KOTHARI; JAIN, 2003).

O método desenvolvido emprega duas técnicas: k -vizinhos-mais-próximos (*Nearest Neighbour Regression - kNNR*) (GUVENIR; UYSAL, 2000; UYSAL; GUVENIR, 2004) e a regressão por processos Gaussianos (*Gaussian Process Regression - GPR*) (RASMUSSEN; NICKISCH, 2010). Problemas de minimização podem ser tratados de maneira semelhante,

alterando o sinal dos valores de y , isto é, substituindo y por seu oposto $-y$. Nesta tese, assume-se a busca para o máximo valor de y .

Para que seja possível a utilização dessas técnicas, é necessário que um pequeno subconjunto de objetos seja definido. Desta forma, inicia-se pela seleção dos n_0 objetos dentre os objetos (\mathbf{x} -vetores) disponíveis, sem utilizar qualquer informação relativa aos valores de y correspondentes, pois em experimentos reais não as temos. Essa seleção inicial é realizada pelo método ODoE. Para efeito de comparação, essa escolha inicial também foi realizada utilizando-se o algoritmo Kennard-Stone (KS).

O KS é um método clássico para seleção de objetos de uma maneira quase uniforme em função das distâncias Euclidianas entre cada par de \mathbf{x} -vetores. O algoritmo KS não se destina especificamente à otimização, uma vez que favorece a exploração de um espaço de busca de forma global.

Vale destacar que tanto o ODoE, quanto o KS, podem selecionar qualquer um dos objetos disponíveis na base de dados, no entanto, somente n_0 objetos serão utilizados para a experimentação inicial.

Assim, após a seleção inicial, são realizados experimentos para que sejam encontrados os valores de propriedade y desejada, ou seja, em uma situação real, os n_0 objetos são submetidos ao método experimental para definição da propriedade de interesse.

Os n_0 objetos e os y associados a cada um desses objetos são utilizadas pelo algoritmo de busca ativa para estimar \hat{y} e, desta forma, encontrar o próximo objeto com o qual pretende-se realizar o experimento. O algoritmo de busca ativa progride de forma sequencial, a partir desse pequeno subconjunto de objetos com valores y conhecidos. Em cada iteração, é utilizada uma técnica de regressão para obter valores estimados para os objetos com valores y desconhecidos. O objeto com melhor valor estimado \hat{y} é selecionado como um candidato provável para otimizar a propriedade desejada.

O melhor valor \hat{y} corresponde ao menor valor em problemas de minimização ou ao maior valor em problemas de maximização. Este objeto candidato é então submetido ao procedimento de análise direta para avaliação da propriedade y .

4.5 Algoritmo de busca inicial

Para que seja possível a utilização da busca ativa, é necessário que um pequeno subconjunto n_0 de objetos seja definido. Esses objetos são submetidos à experimentação para identificação da propriedade de interesse.

4.5.1 Planejamento ótimo de experimentos

O planejamento de superfície de resposta, como o planejamento de composto central ou o planejamento de Box-Behnken, são amplamente usados porque são planejamentos bastante gerais e flexíveis (MONTGOMERY, 2013). No entanto, ocasionalmente, pode-se encontrar uma situação em que um projeto de superfície de resposta não oferece uma boa solução; neste caso, o planejamento ótimo de experimentos pode ser considerado (MONTGOMERY, 2013).

Neste trabalho foi utilizada a técnica de planejamento ótimo de experimentos para seleção dos n_o objeto iniciais. Convém ressaltar que, neste texto, as componentes do objeto são os fatores do ODoE. Para a estimativa das propriedades de interesse é necessário escolher o modelo de ajuste para o ODoE que pode ser linear ou quadrático.

A regressão linear contendo somente uma variável independente é chamada de regressão linear simples. A regressão linear contendo mais que uma variável independente é referida como uma regressão linear múltipla (CORDEIRO, 1993). O modelo de regressão linear é dado pela Equação 4:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \quad (4)$$

na qual, y_i representa os valores observados da variável dependente; $\beta_0, \beta_1, \beta_2$ são parâmetros a serem estimados (são desconhecidos), chamados de coeficientes de regressão; x_1, x_2 representam os valores das variáveis independentes; e o termo ε corresponde ao erro experimental relacionados com os valores observados y , que, em geral, são considerados independentes e normalmente distribuídos com média zero e variância constante

O modelo de regressão quadrática ou de regressão polinomial quadrática, com uma variável independente, é dado pela Equação 5:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \quad (5)$$

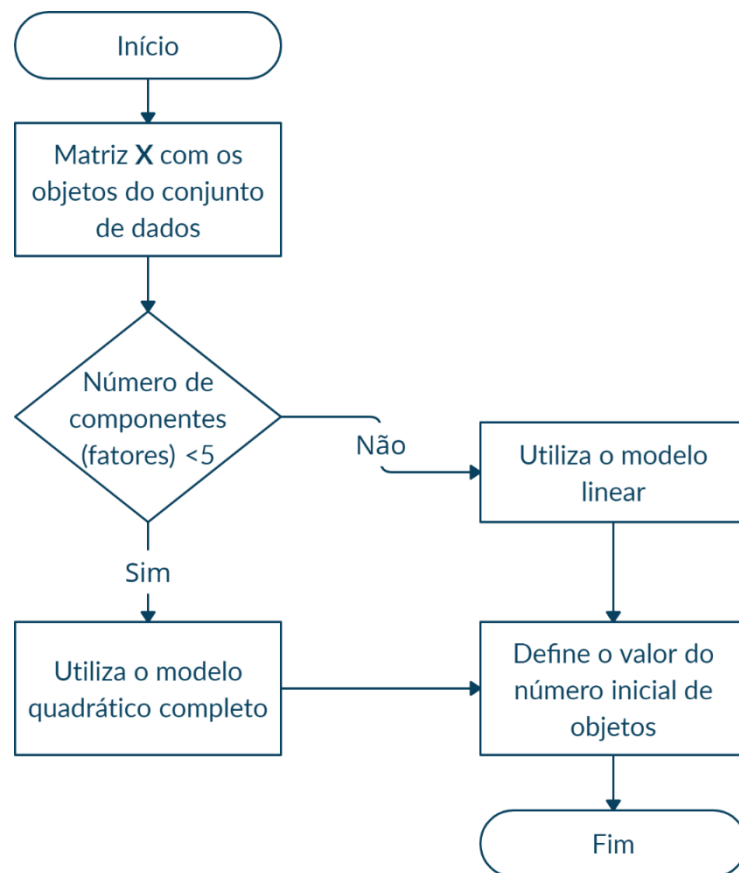
Quanto maior for o número de componentes do modelo, maior sua complexidade, por essa razão, para definir seu modelo de ajuste do ODoE, o número de componentes dos objetos é analisado conforme o fluxograma descrito na Figura 4.4.

Por exemplo, o modelo quadrático completo para duas componentes de um objeto é dado pela Equação 6. Nesse exemplo, a quantidade de objetos iniciais n_o é o mesmo do número de parâmetros a serem estimados (incógnitas). Para duas componentes, é utilizado o modelo quadrático completo.

$$\hat{y}(c_1, c_2) = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \hat{\beta}_{11} c_1^2 + \hat{\beta}_{22} c_2^2 + \hat{\beta}_{12} c_1 c_2, \quad (6)$$

na qual, \hat{y} é o valor estimado da propriedade desejada; c_1 e c_2 são as componentes do objeto; e $\hat{\beta}$ representa o estimador dos parâmetros do modelo, que podem ser obtidos pelo método dos mínimos quadrados.

Figura 4.4. Fluxograma para escolha do modelo de ajuste do ODoE.



Fonte: A autora.

O algoritmo ODoE recebe uma matriz \mathbf{X} contendo os objetos do conjunto de dados e retorna i_{SEL} que contém n_0 índices que indexam os objetos de \mathbf{X} , definidos para o experimento inicial.

4.5.2 Algoritmo de Kennard-Stone

No algoritmo KS, \mathbf{X} é a matriz com N objetos do conjunto de dados, n é o número de objetos e n_0 é o número de objetos iniciais. Assume-se que o valor de $n_0 \geq 2$. Sejam i_{SEL} e $i_{\overline{SEL}}$ os conjuntos de índices dos objetos selecionados e dos objetos ainda disponíveis para

seleção, respectivamente. A descrição formal desse algoritmo é apresentada no Quadro 4.1, na forma de uma sequência de passos.

O algoritmo recebe \mathbf{X} e n_o e retorna i_{SEL} que contém n_o índices dos objetos selecionados, $i_{\overline{SEL}}$ que contém os índices dos objetos não selecionados e a matriz de distâncias Euclidianas (\mathbf{D}).

Inicialmente, nenhum objeto foi selecionado e, portanto, $i_{SEL} = \emptyset$ e $i_{\overline{SEL}} = \{1, 2, \dots, N\}$. É calculada a distância Euclidiana entre cada par dos vetores $\mathbf{x}_i, \mathbf{x}_j$ para $i, j = 1, 2, \dots, N$ em que $i \neq j$.

Em seguida, é feita a seleção dos dois primeiros objetos, considerando que ind_1 e ind_2 indicam os objetos separados pela maior distância.

Depois, é feita a atualização dos vetores que contém os índices dos objetos, movendo ind_1 e ind_2 de $i_{\overline{SEL}}$ para i_{SEL} . Cada objeto subsequente é escolhido segundo o critério de máxima distância mínima, com o objetivo de evitar a seleção de objetos próximos.

Quadro 4.1. Algoritmo de Kennard-Stone

Algoritmo 1 Kennard-Stone

Input: \mathbf{X}, n_o

Output: i_{SEL} (índices dos objetos selecionados)

01: $i_{SEL} = \emptyset$

02: $i_{\overline{SEL}} = \{1, 2, \dots, N\}$

03: **for** $i = 1, \dots, N$ **do**

04: **for** $j = 1, \dots, N$ **do**

05: $\mathbf{D}(i, j) = |\mathbf{x}_i - \mathbf{x}_j|$ // distância Euclidiana

06: **end for**

07: **end for**

08: // índice dos dois objetos separados pela maior distância

09: $(ind_1, ind_2) = \arg_{i, j \in \{1, 2, \dots, N\}} \max \mathbf{D}(i, j)$

10: $i_{SEL} = \{ind_1, ind_2\}$

11: $i_{\overline{SEL}} = i_{\overline{SEL}} \setminus \{ind_1, ind_2\}$

12: $n = 2$

13: **do**

14: **for** $i = 1, \dots, n$ **do**

15: $ind_{n+1} = \arg_{ind \in i_{\overline{SEL}}} \max [\min \mathbf{D}(ind, i)_{i \in i_{SEL}}]$

16: $i_{SEL} = i_{SEL} \cup ind_{n+1}$

17: $i_{\overline{SEL}} = i_{\overline{SEL}} \setminus ind_{n+1}$

18: **end for**

19: $n = n + 1$

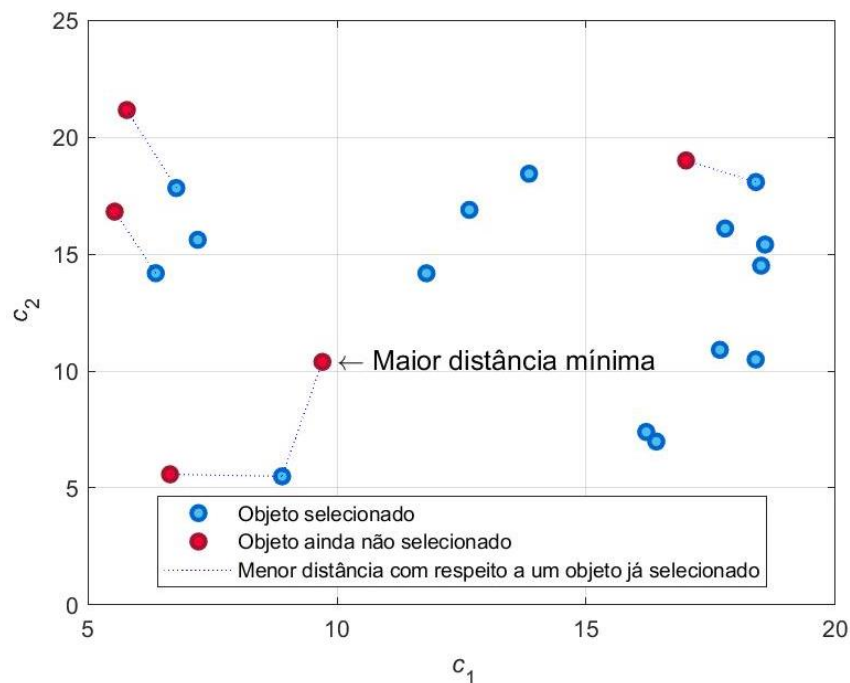
20: **while** $(n \leq n_o)$

21: **return** $i_{SEL}, i_{\overline{SEL}}, \mathbf{D}$

Fonte: A autora.

É feita a atualização dos vetores que contém os índices dos objetos, movendo ind_{n+1} de $i_{\overline{SEL}}$ para i_{SEL} . O método é ilustrado pela Figura 4.5, na qual é mostrada a maior distância mínima entre objetos selecionados (em azul) e ainda disponíveis (em vermelho). O índice do objeto não selecionado, com maior distância mínima, será transferido para o conjunto de índices já selecionados i_{SEL} . Em seguida, o valor de n é incrementado enquanto $n \leq n_0$. Desta forma, os objetos restantes do conjunto de dados são selecionados de acordo com o procedimento máximo-mínimo. Este processo continua até que o número desejado de objetos tenha sido selecionado.

Figura 4.5. Objeto com a maior distância mínima do objeto já selecionado é indicado pela seta



Fonte: A autora.

Após a descrição dos métodos iniciais de busca, na próxima seção serão discutidos os métodos e busca ativa propostos.

4.6 Busca ativa

Foram desenvolvidas duas técnicas de busca ativa para identificação de propriedades ótimas. Essas técnicas são descritas a seguir.

4.6.1 Método de busca ativa utilizando k NNR

O k NNR é um método de aprendizado de máquina robusto, simples de implementar e computacionalmente não custoso (DURBIN *et al.*, 2021; WANG; TAN, 2016). O número de vizinhos é especificado pelo usuário e os k -vizinhos-mais-próximos são definidos pelos k pontos com a menor distância Euclidiana no espaço de estado.

O método proposto envolve um algoritmo que seleciona um novo objeto a cada iteração. Assume-se que os valores de y foram determinados para os n_0 objetos já selecionados pelo algoritmo de seleção inicial (ODoE ou KS).

A técnica de regressão é então empregada para estimar valores de y para os objetos ainda não selecionados. O(s) objeto(s) com maior valor estimado para y é (são) então selecionado(s) para experimentação. É determinada a propriedade de interesse desse(s) objeto(s) (seja por experimentação ou simulação). Esse processo é repetido para cada objeto com índice ainda não selecionado até que todos os índices sejam selecionados ou até que se chegue a um número máximo de experimentos (n_{max}), definido por um especialista da área.

O algoritmo k NNR recebe a matriz \mathbf{X} , contendo os N objetos do conjunto de dados; o vetor y , com os valores associados ao conjunto dos objetos selecionados; a matriz \mathbf{D} , contendo as distâncias Euclidianas; \mathbf{i}_{SEL} que contém n_0 índices dos objetos selecionados pelo algoritmo de inicialização (ODoE ou KS) e que possuem a propriedade de interesse correspondente a cada objeto; $\mathbf{i}_{\overline{SEL}}$ que armazenam os índices dos objetos não selecionados, n_0 e n_{max} , que são o número inicial e o número máximo de objetos, respectivamente. E retorna \mathbf{i}_{SEL} que contém n_{max} objetos selecionados.

Os dois parâmetros que precisam ser escolhidos pelo experimentador são o valor de k , correspondente ao k -vizinhos-mais-próximos, e número total de objetos dado por $n_{max} \in [n_0, N]$. No algoritmo, $N_K(k, \mathbf{i}_{SEL})$ é o conjunto dos índices do k -vizinhos-mais-próximos de um objeto não selecionado \mathbf{x}_j , sendo que esses devem estar contidos em \mathbf{i}_{SEL} . Nesse caso, foi considerado $k = n$ para cada interação do algoritmo. Esta alternativa tem a vantagem de eliminar a necessidade de escolha de um número fixo para o k -vizinhos-mais-próximos.

A estimativa dos valores empregando o k NNR se dá conforme a Equação 7:

$$\hat{y}_j = \frac{\sum_{ind \in N_k(k, \mathbf{i}_{SEL})} \left\| \frac{y_{ind}}{\|\mathbf{x}_{ind} - \mathbf{x}_j\|} \right\|}{\sum_{ind \in N_k(k, \mathbf{i}_{SEL})} \left\| \frac{1}{\|\mathbf{x}_{ind} - \mathbf{x}_j\|} \right\|}, j \in \mathbf{i}_{\overline{SEL}} \quad (7)$$

Note que, de acordo com a Equação 4, as estimativas dos valores de y são baseadas exclusivamente na distância Euclidiana, e podem ser calculadas mesmo que o número de componentes seja maior que o número de objetos já analisados. Essa é uma vantagem sobre os métodos de regressão linear múltipla, que normalmente requerem o uso de técnicas de seleção de componentes quando o número de objetos é menor do que o número de componentes (MATTA *et al.*, 2016).

Após o cálculo do valor estimado \hat{y} , é feita a seleção do próximo objeto que contém o maior valor estimado, conforme Equação 8:

$$ind_{n+1} = \arg \max_{j \in \mathbf{i}_{\overline{SEL}}} \hat{y} \quad (8)$$

E assim é feita a atualização dos vetores que contém os índices dos objetos, movendo ind_{n+1} de $\mathbf{i}_{\overline{SEL}}$ para \mathbf{i}_{SEL} . Em seguida, o valor de n é incrementado enquanto $n \leq n_{max}$. A descrição formal desse algoritmo é apresentada no Quadro 4.2, na forma de uma sequência de passos.

Quadro 4.2. Algoritmo de busca ativa utilizando a técnica k NNR

Algoritmo 2 Busca ativa utilizando k NNR

Input: $\mathbf{X}, \mathbf{y}, \mathbf{D}, \mathbf{i}_{SEL}, \mathbf{i}_{\overline{SEL}}, n_0, n_{max}$

Output: \mathbf{i}_{SEL} (índices dos objetos selecionados)

01: $n = n_0$

02: **do**

03: $k = n$ // k vizinhos mais próximos

04: **for** $j = 1, \dots, N - n$ **do** // j é o j -ésimo valor do índice $\mathbf{i}_{\overline{SEL}}$

$$06: \quad \hat{y}_j = \frac{\sum_{ind \in N_k(k, \mathbf{i}_{SEL})} \left\| \frac{y_{ind}}{D(ind, ind_j)} \right\|}{\sum_{ind \in N_k(k, \mathbf{i}_{SEL})} \left\| \frac{1}{D(ind, ind_j)} \right\|}, \quad ind_j \in \mathbf{i}_{\overline{SEL}}$$

07: **end for**

08: $ind_{n+1} = \arg_{ind \in \mathbf{i}_{\overline{SEL}}} \max \hat{y}$

09: $\mathbf{i}_{SEL} = \mathbf{i}_{SEL} \cup ind_{n+1}$

10: $\mathbf{i}_{\overline{SEL}} = \mathbf{i}_{\overline{SEL}} \setminus ind_{n+1}$

11: $n = n + 1$

12: **while** ($n \leq n_{max}$)

13: return \mathbf{i}_{SEL}

Fonte: A autora.

Nos testes realizados, com dados simulados e reais, o valor de n_0 é determinado pelo Minitab® (Figura 3.2). Utilizar menos objetos iniciais do que definido pelo planejamento de experimentos não foi apropriado, pois não havia informações suficientes para o procedimento de busca ativa. Por outro lado, o uso de mais objetos iniciais escaparia à finalidade do método

proposto, destinado a utilizar poucos objetos para experimentação. No entanto, a escolha do n_{max} dependerá do tempo e recursos disponíveis para a análise dos objetos ou da utilização de um critério de parada. De fato, ao analisar um número maior de objetos, aumenta a possibilidade de atingir o valor ótimo da propriedade y , contudo, aumenta também o custo do experimento. Convém ressaltar que a experimentação ou simulação é a avaliação da propriedade que se quer maximizar ou minimizar.

4.6.2 Método de busca ativa utilizando GPR

Outra técnica de busca ativa utilizada foi a GPR, essa é uma das abordagens Bayesianas de aprendizado ativo de máquinas baseada em um método particularmente eficaz para definir a distribuição de probabilidade prévia sobre uma função latente (RASMUSSEN, WILLIAMS, 2006). Uma distribuição Gaussiana é uma distribuição de probabilidade que descreve as variáveis aleatórias como escalares ou vetores. Pode ser especificada por uma média e uma covariância, de acordo com a Equação 9:

$$\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2), \quad (9)$$

na qual \mathbf{x} é um vetor aleatório, μ é a média e σ^2 é a variância.

Um processo Gaussiano pode ser considerado uma generalização da distribuição de probabilidade Gaussiana. Do ponto de vista do espaço de funções, um processo gaussiano é uma coleção de variáveis aleatórias, qualquer número finito que possua uma distribuição gaussiana conjunta. O processo Gaussiano é um processo estocástico no qual cada subconjunto finito (com variáveis aleatórias) tem uma distribuição multivariada normal (HERFURTH, 2020). Ou seja, para um conjunto \mathbf{X} , um processo estocástico de valor real $\{f(\mathbf{x}), \mathbf{x} \in \mathbf{X}\}$ é um processo Gaussiano se, para qualquer subconjunto $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \in \mathbf{X}$, $f(\mathbf{x})$ têm uma distribuição gaussiana conjunta, descrito por sua função média μ e sua função de covariância cov .

Vamos assumir que $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, 2, \dots, N\}$ representa o conjunto de treinamento do modelo Gaussiano e N representa o número total de objetos. Os vetores $\mathbf{x}_i \in \mathbf{R}^N$ consistem nos objetos da base de dados. Os valores de y representam a propriedade de interesse. $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{SEL}$ representa a matriz de treinamento, $\mathbf{y} = \{y_i\}_{i=1}^{SEL}$ representa o valor de saída do vetor \mathbf{x}_i . O processo Gaussiano pode ser entendido como uma distribuição probabilística sobre os valores da função, denotado pela Equação 10:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), \text{cov}(\mathbf{x}, \mathbf{x}^*)), \quad (10)$$

na qual $f(\mathbf{x})$ é definida por sua média $\mu(\mathbf{x})$ e sua covariância $\text{cov}(\mathbf{x}, \mathbf{x}^*)$ e \mathbf{x}^* representa um objeto ainda não selecionado.

As observações sempre vêm com algum ruído ou imprecisão, desta forma, é melhor assumir um ruído ε , com variância σ_ε^2 . Nesse caso, a saída y de uma função f com a entrada \mathbf{x} pode ser expressa como na Equação 11:

$$y = f(\mathbf{x}) + \varepsilon, \text{ na qual } \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2). \quad (11)$$

A função de covariância $\text{cov}(\mathbf{x}, \mathbf{x}^*)$, também chamada de *kernel* do modelo GPR, possibilita o cálculo de uma distribuição preditiva gaussiana para qualquer \mathbf{x}^* . Uma função de covariância comumente utilizada é a função quadrática exponencial (DENG *et al.*, 2020; HERFURTH, 2020), conforme Equação 12:

$$\text{cov}(\mathbf{x}, \mathbf{x}^*, l) = \sigma^2 \exp\left(\frac{-|\mathbf{x} - \mathbf{x}^*|^2}{2l^2}\right), \quad (12)$$

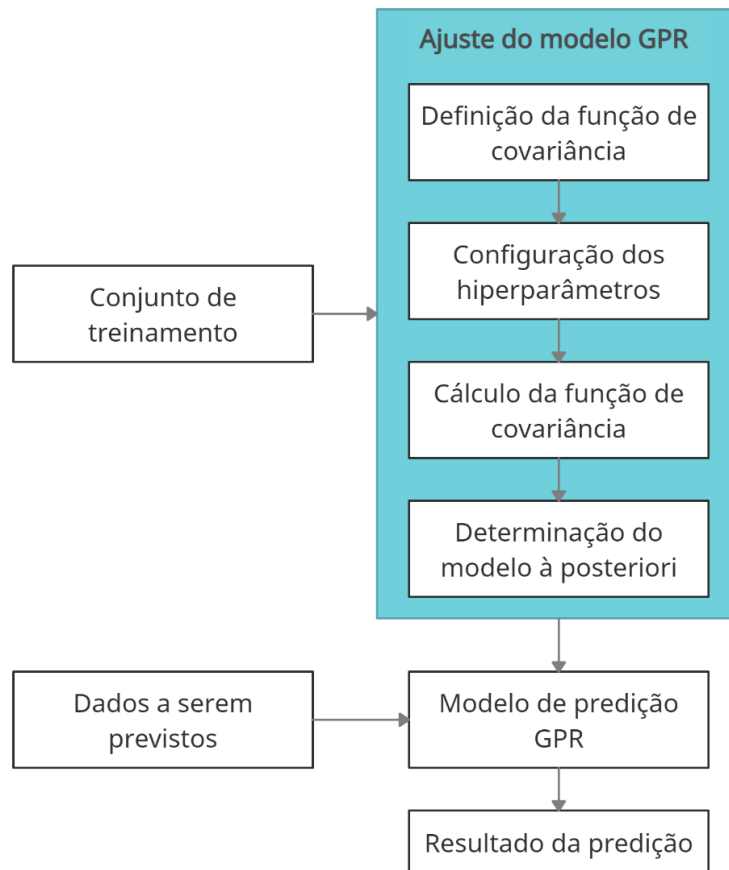
na qual σ e l são hiperparâmetros: σ representa amplitude da função de covariância (amplitude do *kernel*) e l representa a escala de comprimento da medida de distância (DENG *et al.*, 2020).

Um alto valor de l informa que existe uma forte correlação entre dois pontos, separados pela distância $\mathbf{d} = |\mathbf{x} - \mathbf{x}^*|$. Isso significa que a covariância é praticamente independente da característica (do objeto), e pode ser removido da inferência (DENG *et al.*, 2020). Para entradas com vários objetos, especificar l como uma constante significa aplicar a medida de distância isotrópica, que pode acelerar o processo de treinamento, enquanto a especificação de l_i para diferentes objetos de entrada (\mathbf{x}_i) alcançam a determinação automática da relevância, o que pode efetivamente aumentar a precisão da previsão (DENG *et al.*, 2020).

A escolha da função de covariância tem profundo impacto no desempenho de um modelo GPR, pois define a proximidade e similaridade entre os objetos (KONG; CHEN; LI, 2018). A escolha apropriada baseia-se na estrutura dos próprios dados (em termos de suavidade e padrões peculiares). Para os estudos realizados utilizou-se diferentes funções de covariância, descritas na próxima seção.

Após o ajuste do modelo de regressão, de acordo com a Figura 4.6, chamado de treinamento, é feita a predição da propriedade desejada (\hat{y}) para os objetos que contém índices não selecionados. O índice do objeto com maior valor de \hat{y} é então adicionado ao conjunto I_{SEL} . Como a busca ativa visa a melhoria da precisão, mantendo-se sensível também aos custos de otimização, após um primeiro estudo notou-se a necessidade de criar um critério de parada para a técnica GPR, pois o custo computacional pode ser muito alto quando o objeto apresenta muitas componentes. O critério foi estabelecido por meio do cálculo dos intervalos de confiança para os valores estimados \hat{y} . Para este estudo, foi considerado o intervalo de confiança de 95% e, se este intervalo de \hat{y} for maior do que os valores de y dos objetos já selecionados, este objeto é selecionado, caso contrário o algoritmo é encerrado.

Figura 4.6. Ajuste e predição do modelo GRP.



Fonte: A autora.

Como no caso anterior, o algoritmo GPR recebe a matriz \mathbf{X} , contendo os N objetos do conjunto de dados; o vetor y , com os valores associados ao conjunto dos objetos selecionados; a matriz \mathbf{D} , contendo as distâncias Euclidianas; i_{SEL} que contém n_o índices dos objetos selecionados pelo algoritmo de inicialização (ODoE ou KS) e que possuem a propriedade de

interesse correspondente a cada objeto; $i_{\overline{SEL}}$ que armazenam os índices dos objetos não selecionados, n_0 e n_{max} , que são o número inicial e o número máximo de objetos, respectivamente. E retorna i_{SEL} que contém n_{max} objetos selecionados. O algoritmo de busca ativa com a implementação da técnica GPR é descrito no Quadro 4.3.

Quadro 4.3. Algoritmo de busca ativa utilizando a técnica GPR

Algoritmo 3 Busca ativa utilizando GPR

Input: $\mathbf{X}, \mathbf{y}, \mathbf{D}, i_{SEL}, i_{\overline{SEL}}, n_0, n_{max}$

Output: i_{SEL} (índices dos objetos selecionados)

01: $n = n_0$

02: **do**

03: $f(\mathbf{x}_{i_{\overline{SEL}}}) \sim GP(\mu(\mathbf{x}_{i_{SEL}}), cov(\mathbf{x}_{i_{SEL}}, \mathbf{x}_{i_{SEL}^*}))$ // ajuste do modelo de regressão

04: $\hat{\mathbf{y}} = f(\mathbf{x}_{i_{\overline{SEL}}}) + \varepsilon$ // resultado da predição

05: $ind_{n+1} = arg_{ind \in i_{\overline{SEL}}} \max \hat{\mathbf{y}}$

06: **if** $var(\hat{\mathbf{y}}(ind_{n+1})) > (max \mathbf{y}_{i_{SEL}})$

07: $i_{SEL} = i_{SEL} \cup ind_{n+1}$

08: $i_{\overline{SEL}} = i_{\overline{SEL}} \setminus ind_{n+1}$

09: **else**

10: $n = n_{max}$

11: **end if**

12: $n = n + 1$

13: **while** $(n \leq n_{max})$

14: return i_{SEL}

Fonte: A autora.

Após a explanação do método de busca ativa com a técnica GPR, na próxima seção são descritas as funções de covariância utilizadas.

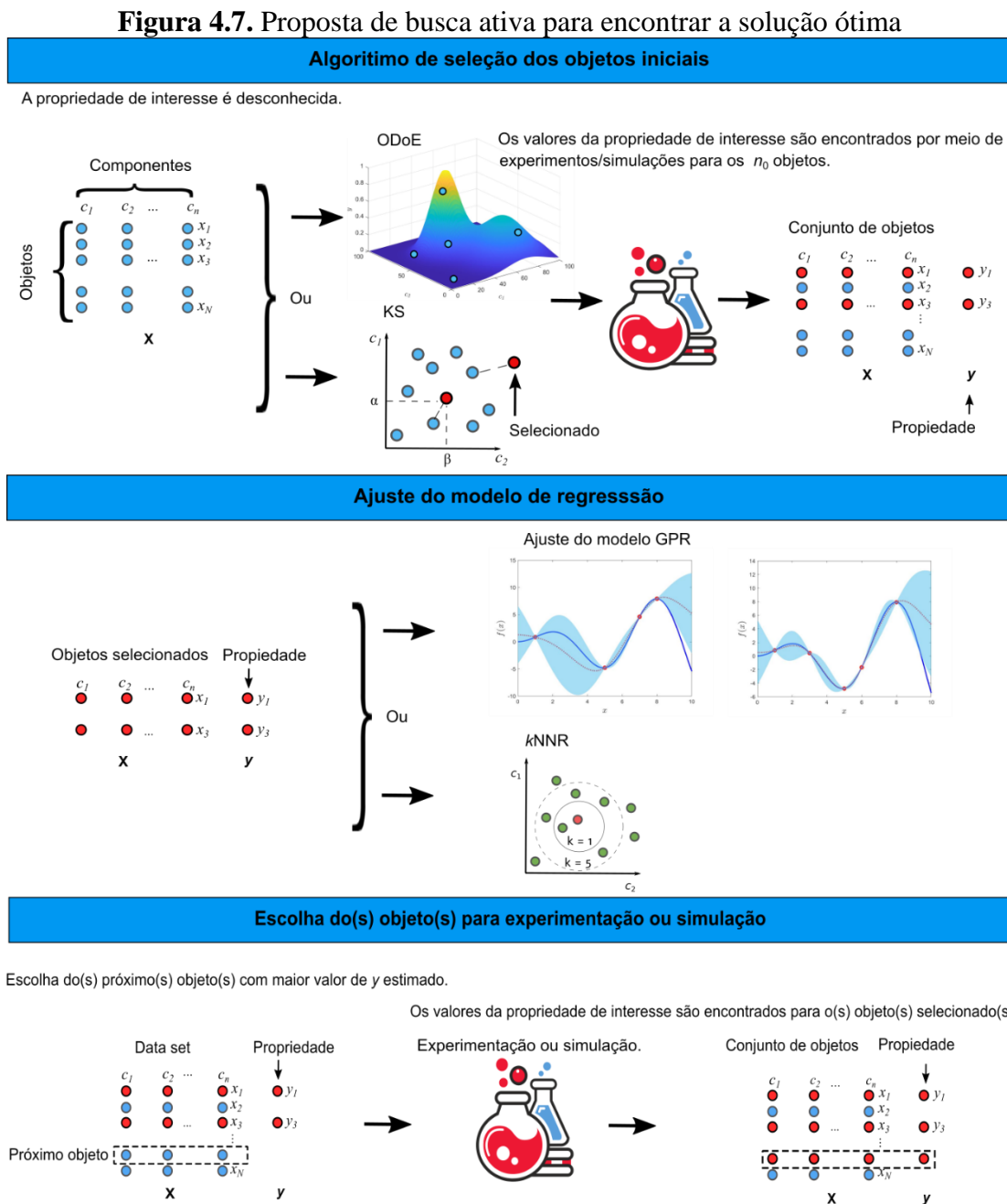
4.6.3 Definição das funções de covariância para a regressão por processos Gaussianos

Vale destacar que, para reprodução de estudos futuros, no primeiro exemplo simulado, adotou-se a função de covariância (*kernel*) exponencial quadrada (*Automatic Relevance Determination* - ARD) para a regressão por processos Gaussianos, implementada pela função MatLab® com o nome de “ardsquaredexponential”. A função de covariância para o segundo exemplo é ARD *Matern* 5/2, implementada pela função MatLab® com o nome de “ardmatern52”. A função de covariância para a base de dados QSAR Medicamentos é ARD exponencial, implementada pela função MatLab® com o nome de “ardexponential” e para a base de dados QSAR Toxicidade é a ARD exponencial, com o nome de “ardmatern32”, no MatLab®. Para a aplicação de microrredes ilhadas, a função de covariância é a exponencial, implementada pela função MatLab® com o nome de “exponential”. A escolha apropriada da

função de covariância baseia-se na estrutura dos próprios dados (em termos de suavidade e padrões peculiares). No MatLab® há um algoritmo que busca ajustar essa função aos dados apresentados para treinamento (*fitrgp*).

4.7 Proposta de busca ativa para encontra a solução ótima

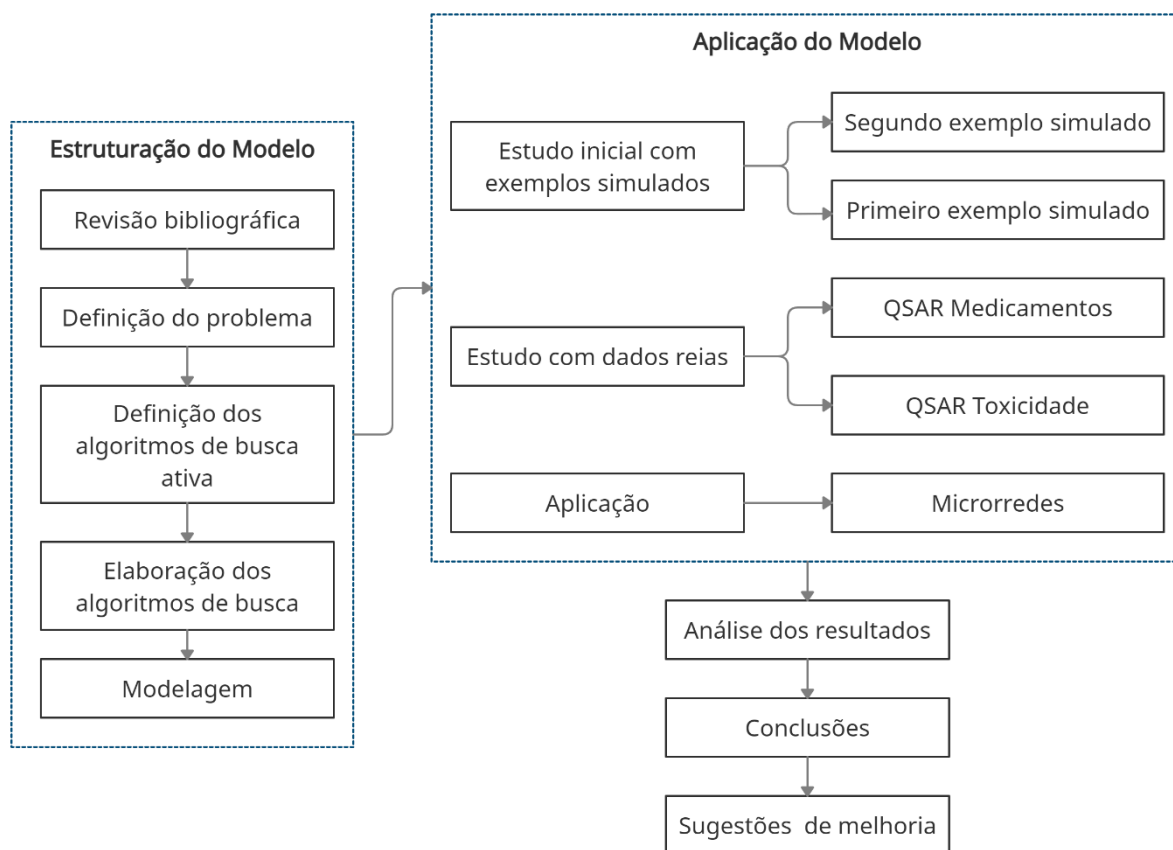
Desta forma, a Figura 4.3, mostra na forma de um resumo gráfico, os principais passos da proposta para solução de problemas desta natureza utilizando a metodologia de busca ativa.



Fonte: A autora.

A Figura 4.8 apresenta, de forma resumida, os procedimentos metodológicos adotados: primeiro, foi feita uma revisão bibliográfica sobre o assunto abordado na tese; depois foi definido o problema de pesquisa, foram definidos e implementados os algoritmos de busca ativa e foi feita a modelagem do problema. Foram definidas as bases de dados para os estudos a serem realizados; o modelo foi então aplicado para obtenção e análise dos resultados, por fim foram feitas as conclusões e foram dadas sugestões para trabalhos futuros.

Figura 4.8. Procedimentos metodológicos



Fonte: A autora.

No próximo capítulo são apresentados os resultados por meio de gráficos e tabelas e é feita uma análise da metodologia de busca ativa proposta.

5 Resultados e discussão da validação da metodologia

Neste capítulo são apresentados os resultados e é feita a análise das soluções encontradas para exemplos simulados, para as bases de dados reais e para o sistema de microrredes.

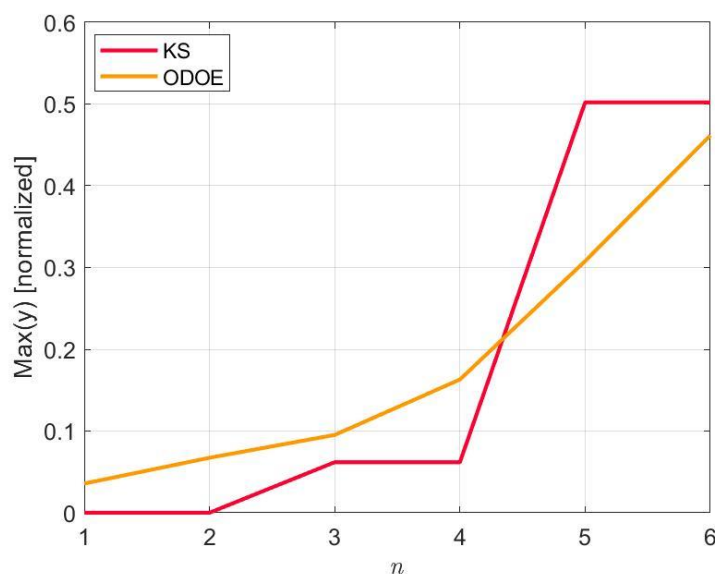
5.1 Primeiro exemplo simulado

Foram feitos três estudos envolvendo o exemplo simulado, associado à função descrita pela Equação 2. Estes estudos são descritos nas próximas seções. As comparações aqui apresentadas são úteis para ilustrar e justificar o mecanismo de busca do método proposto, que enfatiza as regiões associadas aos maiores valores da propriedade de interesse.

5.1.1 Primeiro estudo: seleção dos objetos iniciais

O primeiro estudo fez uma comparação da seleção dos n_0 objetos iniciais entre os métodos de seleção ODoE e KS. A Figura 5.1 mostra o resultado obtido, após a aplicação dos métodos em 100 diferentes subconjuntos contendo 100 objetos cada, gerados por meio de distribuição uniforme. Pode-se notar que esses métodos, após a seleção de 6 objetos iniciais, não encontraram um valor próximo ao ótimo. Isso ocorre porque esses algoritmos buscam representar o espaço amostral nos quais os objetos estão distribuídos, porém não são efetivos para busca de valores ótimos para os problemas apresentados nesta proposta de tese.

Figura 5.1. Relação entre os x -objetos e a propriedade y no exemplo simulado



Fonte: A autora.

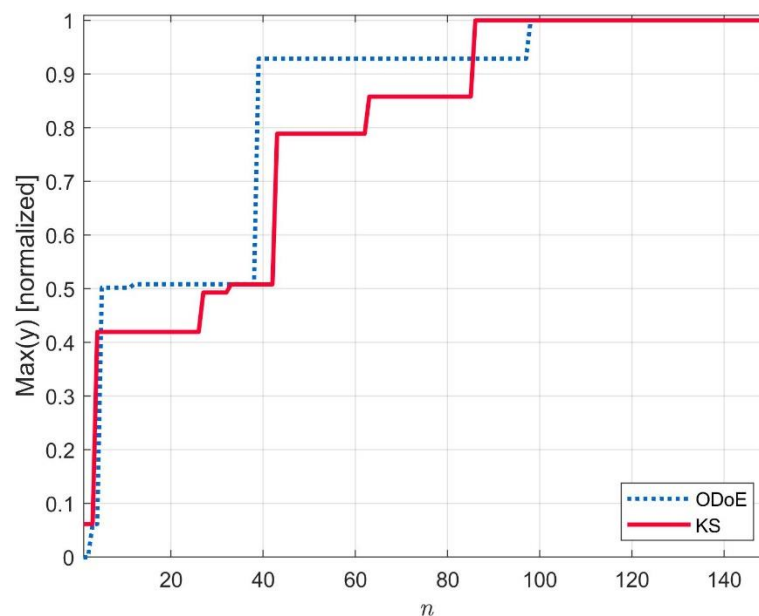
Nota-se que os algoritmos não convergem para um valor ótimo rapidamente, conforme mostrados na Figura 5.2. Nesse estudo, o ponto ótimo, utilizando os algoritmos ODoE e KS foi encontrado na a seleção do 86º objeto e 98º objeto, respectivamente.

Embora ODoE possa levar a resultados sólidos quando combinados com conhecimentos de engenharia adequados, com o fornecimento de orientações quanto à confiabilidade e validade dos resultados, não garante uma cobertura eficiente do espaço amostral (YONDO; ANDRÉS; VALERO, 2018), o que justifica a seleção do ponto ótimo após tantas iterações.

Neste sentido, embora o KS seja uma abordagem sequencial determinística que tenta selecionar amostras uniformemente distribuídas no espaço de predição (RAMIREZ-LOPEZ *et al.*, 2014), a busca realizada pelo algoritmo KS não explora a região associada aos maiores valores de y . Este algoritmo favorece a seleção de objetos que estão distantes uns dos outros, o que impediu a seleção de objetos próximos ao melhor valor da propriedade y já encontrado. Isso justifica porque esse algoritmo não convergiu tão rápido quanto o de busca ativa, como será apresentado nas próximas seções.

Devido a esses fatos, conforme dito anteriormente, tanto o algoritmo ODoE, quanto o KS, foram utilizados como algoritmos iniciais de busca de objetos, nos quais se desconhece o valor da propriedade y de interesse.

Figura 5.2. Resultado da busca com a utilização dos algoritmos ODoE e KS.

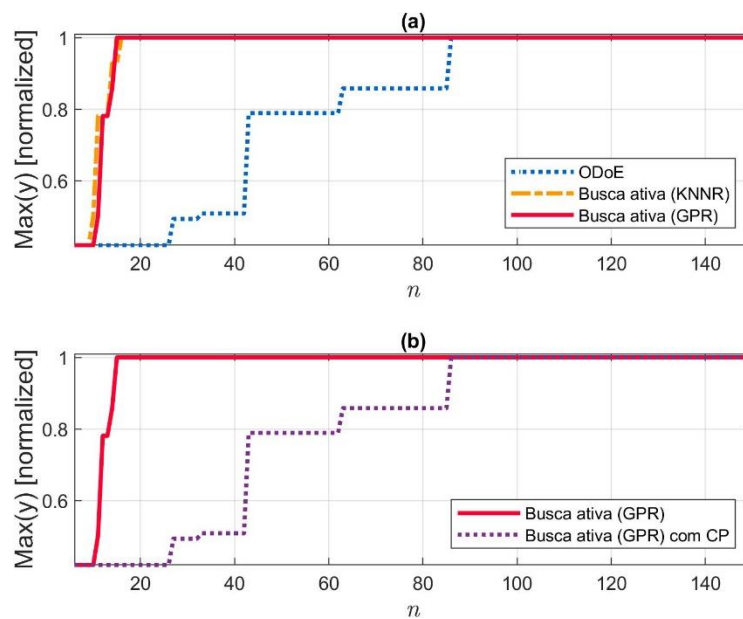


Fonte: A autora.

5.1.2 Segundo estudo: busca ativa

Neste segundo estudo são comparados os resultados de busca ativa por meio dos algoritmos k NNR, GPR e ODoE, sendo a seleção dos objetos iniciais realizada pelo método ODoE. A Figura 6.3 compara os resultados da busca ativa com os obtidos usando o algoritmo ODoE como algoritmo de busca. Como pode ser visto na Figura 5.3 (a), o valor máximo de y é alcançado pelo algoritmo de busca ativa com k NNR e com GPR na avaliação do 15º objeto selecionado. No entanto, ao se utilizar o critério de parada para a técnica GPR, o resultado não foi bom, pois só encontrou o valor ótimo após a avaliação de $n = 86$, conforme Figura 6.3 (b). O algoritmo ODoE encontra o valor ótimo de y , somente após a avaliação de $n = 86$ objetos.

Figura 5.3. Gráfico comparativo da busca ativa utilizando as técnicas k NNR e GPR com inicialização pelo ODoE.



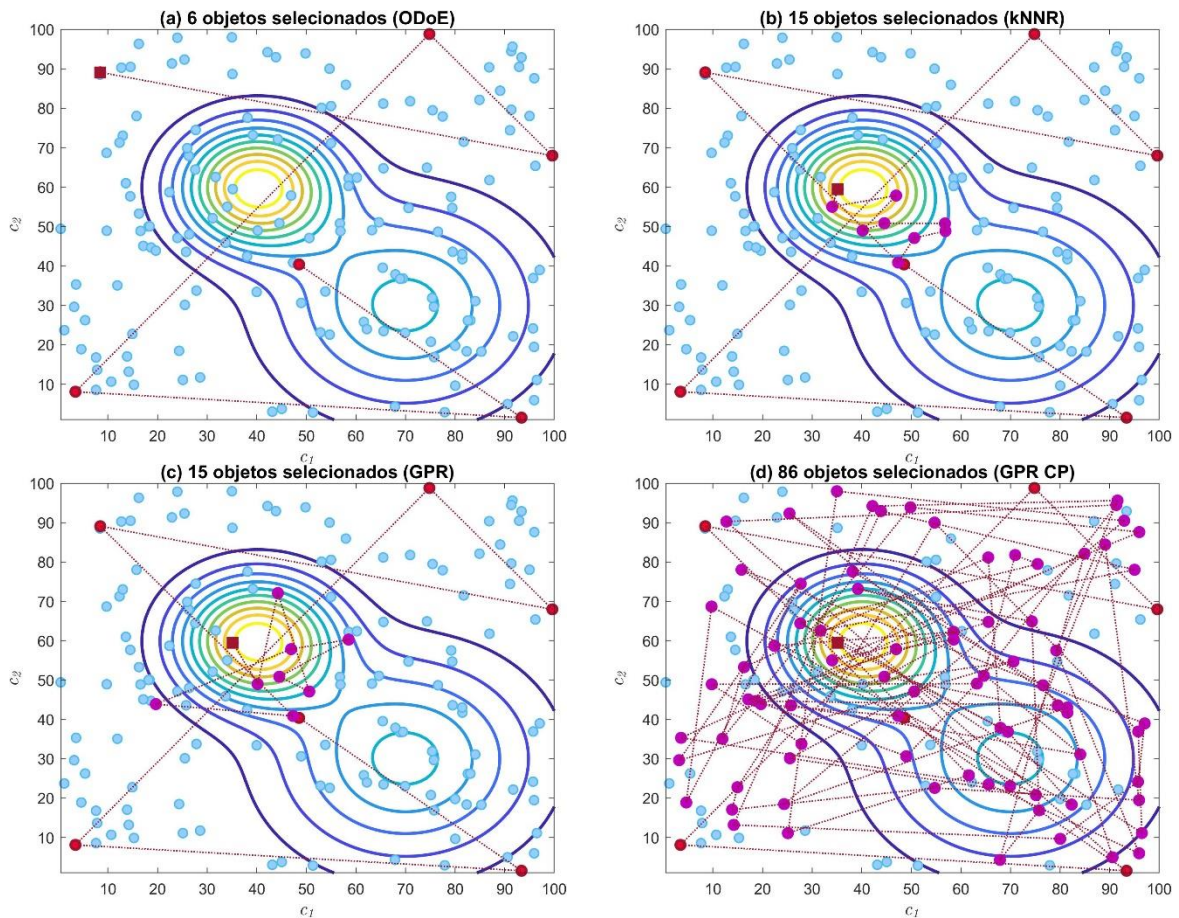
Fonte: A autora.

Nota-se que, com a utilização do algoritmo ODoE para a seleção inicial dos objetos, a busca ativa apresentou os melhores resultados com a utilização das técnicas k NNR e GPR, conforme Figura 6.4. A Figura 5.4 (a) mostra os 6 objetos iniciais, selecionados por ODoE, plotados com marcadores vermelhos conectados por linhas tracejadas. O ponto ótimo utilizando a busca ativa com as técnicas k NNR e GPR, é encontrado no 15º objeto selecionado para avaliação, conforme Figuras 5.4 (b) e (c), respectivamente.

Entretanto, o algoritmo demorou para convergir quando se utilizou o critério de parada para busca ativa com a técnica GPR, de acordo com a Figura 5.4 (d). Uma provável justificativa

para essa demora em achar o ponto ótimo é o fato de, ao se estimar o valor de \hat{y} e selecionar o maior valor para o teste experimental, este valor encontra-se distante do ponto ótimo.

Figura 5.4. Resultados do método de busca ativa no primeiro exemplo simulado: (a) 6 objetos iniciais selecionados pelo algoritmo ODoE; (b) ponto ótimo encontrado no 15º objeto avaliado, utilizando k NNR; (c) ponto ótimo encontrado no 15º objeto avaliado, utilizando GPR; e (d) ponto ótimo encontrado no 86º objeto, utilizando GPR com critério de parada.



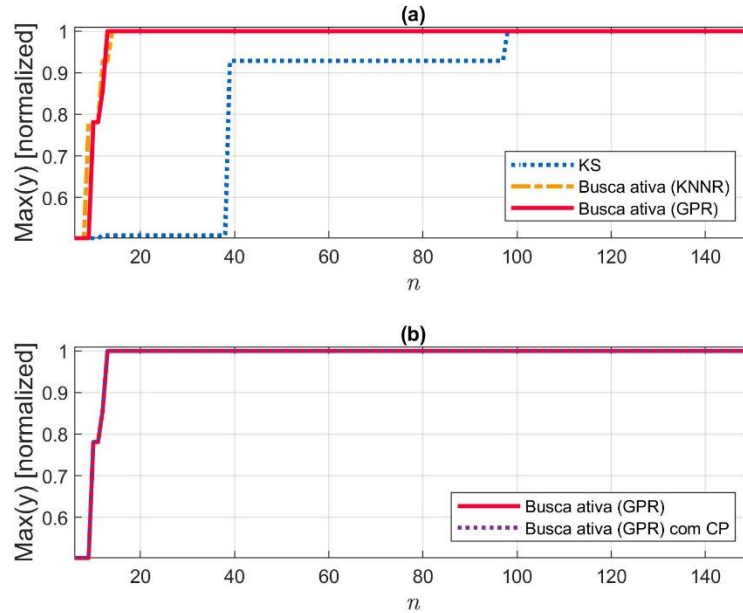
Fonte: A autora.

5.1.3 Terceiro estudo: busca ativa

O terceiro estudo com dados simulados utilizou o algoritmo KS para seleção dos n_0 objetos. A Figura 6.5 compara os resultados da busca ativa com os obtidos usando o algoritmo KS. Os 150 objetos envolvidos neste exemplo são indicados em um gráfico de contorno da função associada aos valores y . Como pode ser visto, o valor máximo de y é alcançado rapidamente pelo algoritmo de busca ativa com k NNR e GPR após a avaliação de $n = 14$ e $n = 13$ objetos, respectivamente, enquanto para o KS encontra o valor ótimo de y , somente após a avaliação de $n = 98$ objetos, de acordo com a Figura 5.5 (a). A Figura 5.5 (b) mostra que a busca

ativa utilizando critério de parada para a técnica GPR apresentou um excelente resultado, se comparado com a Figura 5.4 (b), apresentada na seção anterior.

Figura 5.5. Gráfico comparativo da busca ativa utilizando as técnicas k NNR e GRP com o KS.



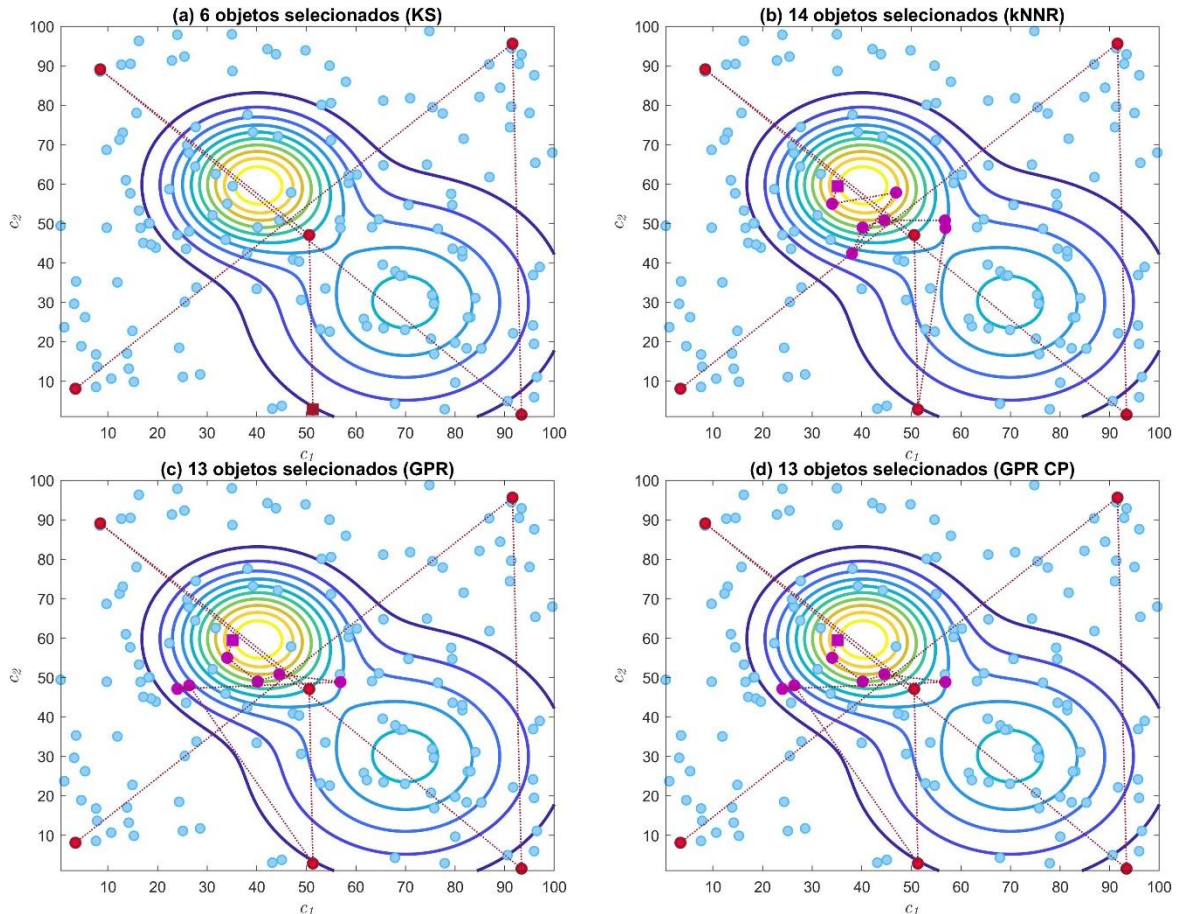
Fonte: A autora.

Similar ao segundo estudo, uma interpretação desses resultados pode ser fornecida com base na Figura 5.6.

Os n_0 objetos selecionados por KS são plotados com marcadores vermelhos conectados por linhas pontilhadas, para este exemplo são 6 objetos iniciais, conforme Figura 5.6 (a). A partir dos n_0 objetos selecionados, a aplicação do método k NNR levou à determinação do maior valor de y , na seleção do 14º objeto, conforme Figura 5.6 (b). As Figuras 5.6 (c) e (d) mostram que na 13ª seleção é encontrado o objeto com maior valor para y , utilizando busca ativa com a técnica GPR, sem e com critério de parada.

A seleção inicial favoreceu uma exploração do espaço \mathbf{x} , com objetos espalhados por todo o plano c_1 - c_2 . Nota-se que, na busca ativa, o algoritmo reduziu o espaço de busca, deixando uma pequena região para se concentrar durante a fase de exploração.

Figura 5.6. Resultados do método de busca ativa no primeiro exemplo simulado: (a) 6 objetos iniciais selecionados pelo algoritmo KS; (b) ponto ótimo após a seleção de 14 objetos utilizando k NNR; (c) ponto ótimo encontrado no 13º objeto avaliado, utilizando GPR; e (d) ponto ótimo encontrado no 13º objeto avaliado, utilizando GPR com critério de parada.



Fonte: A autora.

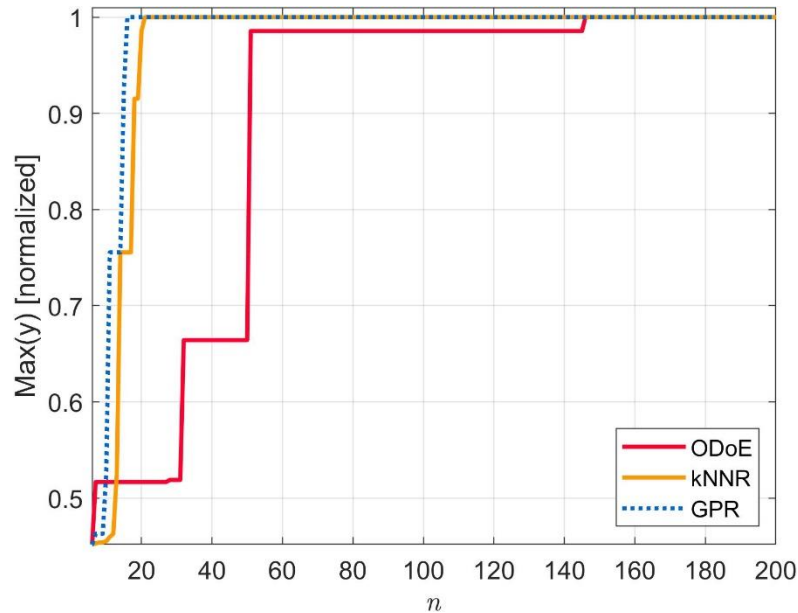
Para dar continuidade ao estudo, outro exemplo simulado foi explorado. Os resultados são apresentados a seguir.

5.2 Segundo exemplo simulado

No segundo exemplo simulado, associado à função descrita pela Equação (3), são comparados os resultados de busca ativa com a busca com o ODoE e com KS. Na Figura 5.7, como pode ser visto, o valor máximo de y é alcançado rapidamente pelo algoritmo de busca ativa após a avaliação de $n = 21$ e $n = 16$ objetos, com as técnicas k NNR e GRP, respectivamente, enquanto para o ODoE encontra o valor ótimo de y somente após a avaliação de $n = 146$ objetos. Na Figura 5.8, o valor máximo de y é alcançado rapidamente pelo algoritmo de busca ativa após

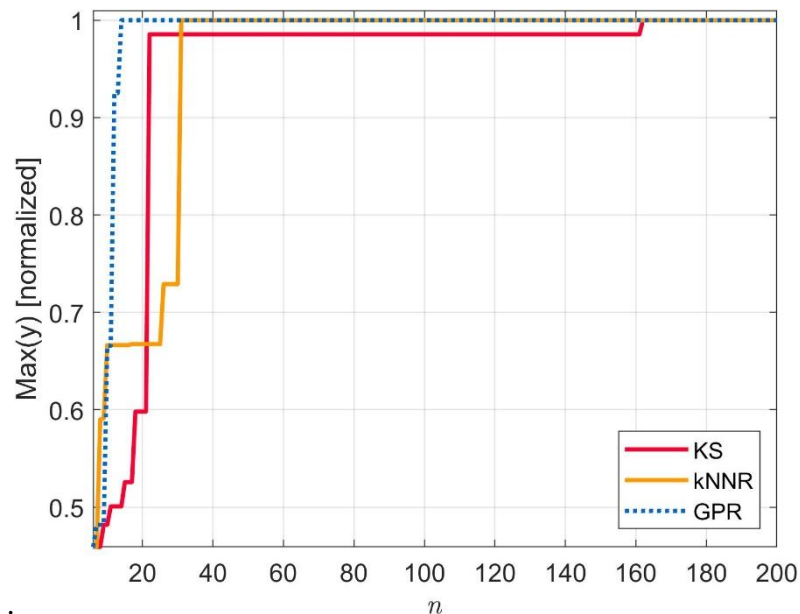
a avaliação de $n = 31$ e $n = 14$ objetos, com a técnicas kNNR e GRP, respectivamente, enquanto para o KS encontra o valor ótimo de y , somente após a avaliação de $n = 162$ objetos.

Figura 5.7. Gráfico comparativo da busca ativa com o ODoE.



Fonte: A autora.

Figura 5.8. Gráfico comparativo da busca ativa com o KS.

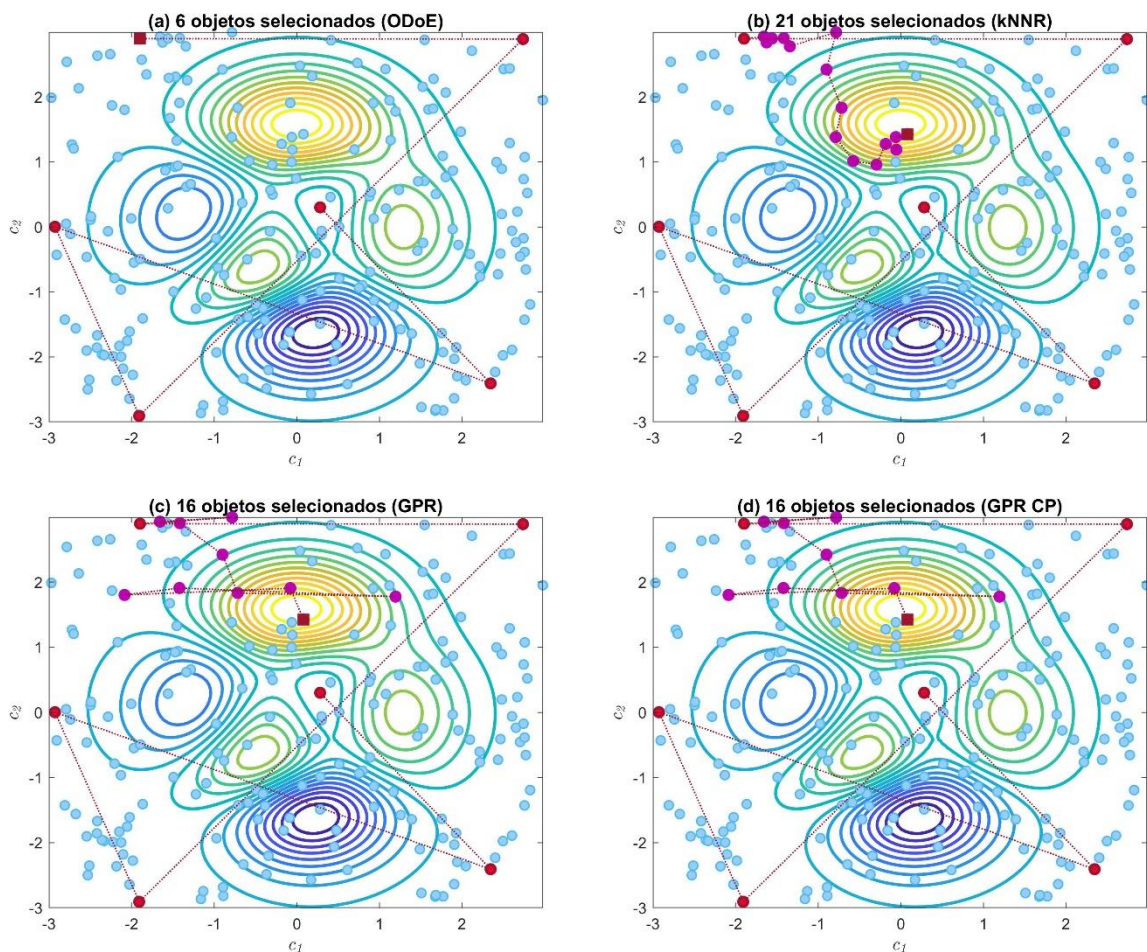


Fonte: A autora.

Como no primeiro exemplo, foi feito um estudo detalhado para compreender o mecanismo do algoritmo de busca ativa em uma função com máximos e mínimos locais. Similar

aos estudos anteriores, uma interpretação desses resultados pode ser fornecida com base nas Figuras 5.9 e 5.10. Os 200 objetos envolvidos neste segundo exemplo simulado são indicados em um gráfico de contorno da função associada aos valores y . Os n_0 objetos selecionados pelo algoritmo ODoE são plotados com marcadores vermelhos conectados por linhas pontilhadas; para este exemplo são 6 objetos iniciais, conforme Figura 5.9 (a). Nota-se que a partir dos n_0 objetos selecionados, a aplicação do método de busca ativa com a técnica k NNR levou à determinação do maior valor de y , após a seleção de 21 objetos, conforme Figura 5.9 (b). As Figuras 5.9 (c) e (d) mostram que, na 16ª seleção, é encontrado o objeto com maior valor para y , utilizando busca ativa com a técnica GPR, sem e com critério de parada.

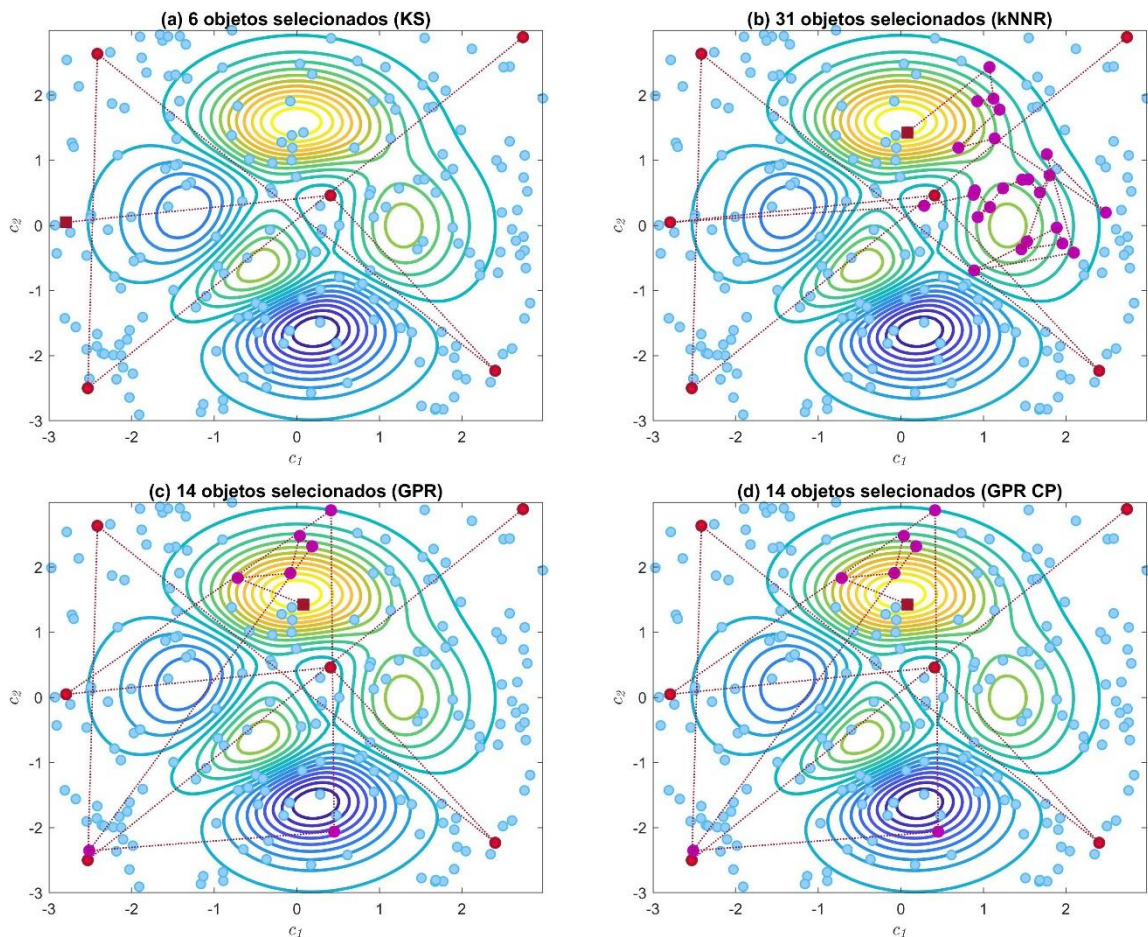
Figura 5.9. Resultados do método de busca ativa para o **segundo** exemplo simulado: (a) 6 objetos iniciais selecionados pelo algoritmo ODoE; (b) ponto ótimo encontrado no 21º objeto avaliado, utilizando k NNR; (c) ponto ótimo encontrado no 16º objeto avaliado, utilizando GPR sem critério de parada e (d) com critério de parada.



Fonte: A autora.

A Figura 5.10 ilustra o uso da técnica de seleção inicial KS, seguida da busca ativa. Os objetos iniciais selecionados pelo algoritmo KS são plotados com marcadores vermelhos conectados por linhas pontilhadas, para este exemplo são 6 objetos iniciais, conforme Figura 5.10 (a). Nota-se que a partir dos n_0 objetos selecionados, a aplicação do método k NNR levou à determinação do maior valor de y na seleção do 31º objeto, conforme Figura 5.10 (b). As Figuras 5.10 (c) e (d) mostram que, na 14ª seleção é encontrado o objeto com maior valor para y , utilizando busca ativa com a técnica GPR, sem e com critério de parada.

Figura 5.10. Resultados do método de busca ativa no exemplo simulado: (a) 6 objetos iniciais selecionados pelo algoritmo KS; (b) ponto ótimo encontrado no 13º objeto avaliado, utilizando k NNR; (c) ponto ótimo encontrado no 14º objeto avaliado, utilizando GPR sem critério de parada; e (d) com critério de parada.



Fonte: A autora.

Pode-se notar que a seleção inicial favoreceu uma exploração do espaço x , com objetos espalhados por todo o plano c_1 - c_2 , tanto com a utilização do algoritmo ODoE, quanto com o algoritmo KS, similar ao primeiro exemplo simulado.

Como no primeiro exemplo, o algoritmo de busca ativa reduziu o espaço de busca, deixando uma pequena região para se concentrar durante a fase de exploração, isso pode ser considerado como uma característica para um bom algoritmo de busca, pois aumenta a capacidade de busca para descobrir potenciais candidatos promissores.

Após estes estudos iniciais com dados simulados, foram realizados estudos com dados reais de QSAR Medicamentos, QSAR toxicidade e foi desenvolvida uma aplicação para determinação de um agrupamento ótimo para um sistema de microrredes ilhadas.

5.3 QSAR Medicamento

A base de dados deste estudo de caso refere-se às relações quantitativas entre a estrutura química e a atividade biológica (QSAR) envolvendo um conjunto de compostos antidepressivos (arilpiperazina) com constantes de inibição (K_i). Neste tipo de problema, as componentes de x são os 14 descritores moleculares dos compostos em consideração que são derivados por meio de cálculos teóricos usando um *software* adequado. Em contraste, a propriedade y é obtida por procedimentos experimentais, como estudos *in vitro* ou *in vivo*.

QSAR é uma ferramenta quimiométrica importante para prever a atividade do composto sem a necessidade de dados experimentais, é amplamente utilizada para descrever quantitativamente as relações entre a estrutura química das moléculas e a atividade biológica por elas desempenhadas, visa a identificação de valores ótimos para determinadas propriedades físico-químicas e, por meio delas, fundamenta-se o planejamento de novas substâncias que possuam perfil terapêutico (ROY *et al.*, 2020).

Em geral, leva cerca de 15 anos e até 1,5 bilhões de dólares para converter um novo composto promissor em um medicamento (DIMASI; GRABOWSKI; HANSEN, 2016). As abordagens e metodologias usadas no projeto de medicamentos foram alteradas ao longo do tempo. No entanto, o estágio inicial desse processo é a utilização de abordagens computacionais, no qual o foco está na redução do número de compostos candidatos à elaboração de medicamentos (KUNDU; PAUL; BANERJEE, 2018). Neste sentido, o algoritmo de busca ativa apresentado nesta tese visa a redução do número de compostos candidatos.

Para validar o método de busca ativa proposto, utilizou-se uma base de dados de QSAR contendo um conjunto de arilpiperazinas. Compostos arilpiperazínicos constituem a classe mais importante de ligantes do receptor 5-HT_{1A}. Esse receptor é um subtipo de serotonina (5-HT) utilizado na regulação do humor e do tratamento da depressão. O receptor 5-HT_{1A} é considerado crítico para tais funções por estar abundantemente localizado nas regiões corticais que estão

implicadas no humor e na emoção, por isso representa um importante alvo de desenvolvimento de agentes terapêuticos para tratar tais disfunções (MATTA *et al.*, 2016; WANG *et al.*, 2019; WEBER *et al.*, 2010).

A função objetivo desse estudo é encontrar o composto com o maior valor de pK_i em cada subconjunto, pois esse é o que possui maiores afinidades pelo receptor 5-HT_{1A}. A Tabela A.1 (Apêndice A) apresenta a estrutura e os valores de propriedades biológicas dos compostos arilpiperazínicos empregados nesta investigação. A propriedade biológica corresponde aos valores experimentais de afinidade para o receptor 5-HT_{1A} indicando termos da constante de inibição K_i (MARTÍNEZ-ESPARZA *et al.*, 2001). No presente estudo, esta propriedade é expressa como $pK_i = -\log K_i$, com valores entre 5,3 e 8,3 (esses valores foram normalizados).

A otimização da geometria de todos os compostos foi, então, realizada usando o método semiempírico Austin Model 1 (AM1) (DEWAR *et al.*, 1985). Para estas estruturas, foram calculados 14 descritores eletrônicos no nível AM1, como mostrado na Quadro 5.1. Todos esses descritores foram assumidos para representar propriedades eletrônicas dos compostos. Após análise da covariância, apresentada em breve, os descritores HL_{GAP} , η e χ não foram utilizados nos cálculos subsequentes, uma vez que possuem informações redundantes. Os valores dos descritores foram auto-escalonados e, desta forma, garantir que as influências relativas de diferentes variáveis sejam independentes em suas unidades.

Quadro 5.1. Descritores eletrônicos e suas definições

Identificador	Descritor	Definição
c_1	$\Delta_f H$	Calor de formação
c_2	α	Polaridade
c_3	E_T	Energia total
c_4	μ	Momento dipolo
c_5	E_{HOMO}	Energia do orbital molecular ocupado mais alto (HOMO)
c_6	E_{LUMO}	Energia do orbital molecular desocupado mais baixo (LUMO)
c_7	QN1	Carga atômica em nitrogênio 1
c_8	QN4	Carga atômica em nitrogênio 4
c_9	QZ	Carga atômica no átomo Z
c_{10}	QC2 _{Ar1}	Carga atômica em carbono 2 do anel Ar1
c_{11}	QS2 _{Ar1}	Carga total em carbono 2 do anel Ar1
c_{12}	HL_{GAP}	Energia entre o HOMO e o LUMO
c_{13}	η	$\eta = \frac{E_{LUMO} - E_{HOMO}}{2}$
c_{14}	χ	$\chi = \frac{E_{HOMO} - E_{LUMO}}{2}$

Fonte: A autora.

Uma análise de componente principal (PCA) usando o *software* Unscrambler® foi empregada a fim de identificar compostos fora do domínio de aplicabilidade. Dessa forma os compostos 20 e 41 (Tabela A.1) apresentaram valores fora dos limites e, portanto, foram removidos do conjunto de dados (MATTA *et al.*, 2016).

Com o objetivo de avaliar o método de busca ativa proposto, os 79 compostos restantes foram utilizados para gerar 100 diferentes subconjuntos, utilizando um procedimento de subamostragem. Cada subconjunto é composto por 50 compostos (objetos) escolhidos de forma aleatória.

Foi feita uma análise da correlação das componentes de \mathbf{X} na busca por reduzir a dimensão do problema (e com isso o número de fatores para o ODoE), conforme Tabela 5.1. Para ilustrar o problema, a Figura 5.11 mostra a correlação entres as componentes de forma gráfica.

Tabela 5.1. Correlação entre as componentes de QSAR Medicamentos.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}
c_1	1,00	-0,36	1,00	-0,30	0,31	-0,23	0,19	0,16	0,07	-0,13	0,04	-0,41	-0,41	0,03
c_2	-0,36	1,00	-0,35	0,17	-0,08	-0,10	0,15	-0,22	-0,03	0,23	-0,21	-0,05	-0,05	0,12
c_3	1,00	-0,35	1,00	-0,29	0,29	-0,25	0,17	0,16	0,05	-0,15	0,05	-0,42	-0,42	0,05
c_4	-0,30	0,17	-0,29	1,00	-0,42	-0,46	-0,24	0,19	0,12	0,01	-0,11	-0,20	-0,20	0,55
c_5	0,31	-0,08	0,29	-0,42	1,00	0,27	0,23	-0,51	0,16	0,12	0,15	-0,33	-0,33	-0,68
c_6	-0,23	-0,10	-0,25	-0,46	0,27	1,00	0,04	-0,32	-0,23	-0,05	0,06	0,82	0,82	-0,89
c_7	0,19	0,15	0,17	-0,24	0,23	0,04	1,00	-0,07	-0,02	0,12	-0,29	-0,10	-0,10	-0,14
c_8	0,16	-0,22	0,16	0,19	-0,51	-0,32	-0,07	1,00	-0,20	0,10	-0,26	-0,01	-0,01	0,48
c_9	0,07	-0,03	0,05	0,12	0,16	-0,23	-0,02	-0,20	1,00	0,17	0,00	-0,32	-0,32	0,10
c_{10}	-0,13	0,23	-0,15	0,01	0,12	-0,05	0,12	0,10	0,17	1,00	-0,78	-0,12	-0,12	-0,02
c_{11}	0,04	-0,21	0,05	-0,11	0,15	0,06	-0,29	-0,26	0,00	-0,78	1,00	-0,03	-0,03	-0,12
c_{12}	-0,41	-0,05	-0,42	-0,20	-0,33	0,82	-0,10	-0,01	-0,32	-0,12	-0,03	1,00	1,00	-0,47
c_{13}	-0,41	-0,05	-0,42	-0,20	-0,33	0,82	-0,10	-0,01	-0,32	-0,12	-0,03	1,00	1,00	-0,47
c_{14}	0,03	0,12	0,05	0,55	-0,68	-0,89	-0,14	0,48	0,10	-0,02	-0,12	-0,47	-0,47	1,00

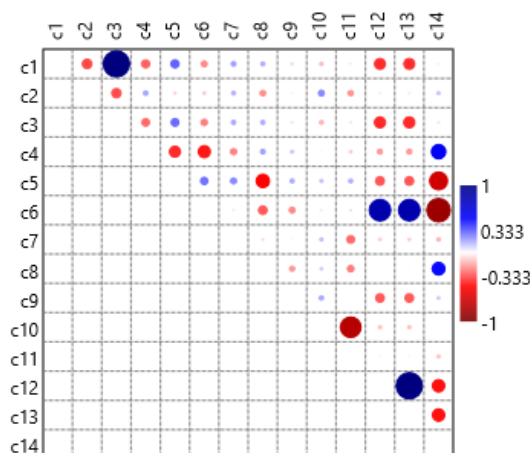
Fonte: A autora.

O coeficiente de correlação mede a direção e o grau de associação linear entre as variáveis. Valores próximos de -1 ou +1 indicam forte associação, portanto essas componentes foram retiradas do vetor \mathbf{x}_i . Foram retirados das amostras as componentes c_3 , c_{12} , c_{13} e c_{14} (destacadas na Tabela 5.1) que possuem alta correlação (círculos maiores da Figura 5.11), sobrando 10 componentes.

Neste problema, foi utilizado o mesmo número de objetos iniciais, tanto utilizando a seleção pelo método ODoE, como para o KS. Para definição desse número foi utilizado o

método fatorial fracionado. A adoção do arranjo fatorial completo inviabiliza o objetivo desse trabalho de selecionar poucos objetos para se chegar à propriedade ótima se o número de componentes é alto, conforme discutido no Capítulo 3. Assim, se chegou ao valor de $n_o = 16$ (objetos iniciais).

Figura 5.11. Correlação entre as componentes do vetor x_i na base QSAR Medicamento.



Fonte: A autora.

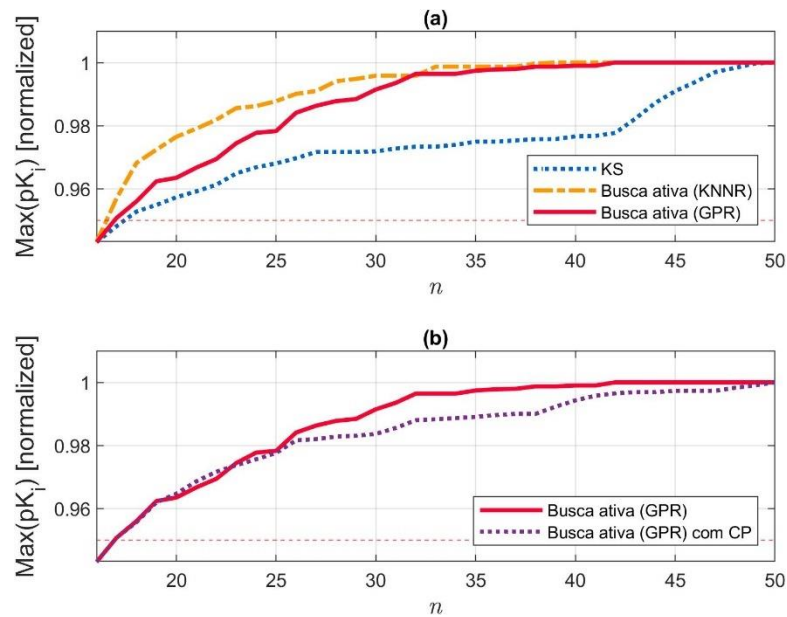
O método de pesquisa ativa proposto foi então aplicado a cada um destes subconjuntos, com o objetivo de encontrar o composto com o maior valor de pK_i em cada subconjunto.

Os resultados mostram que a busca ativa convergiu rapidamente, conforme Figuras 5.12 e 5.13, respectivamente. Valores próximos ao ótimo (acima de 0,95) foram obtidos com menos de 20 objetos selecionados para a maioria dos conjuntos amostrais, o que mostra a eficiência do método proposto com objetos iniciais selecionamos por meio do KS e do ODoE. As Figuras 5.12 (a) e 5.13 (a) mostram comparação entre a busca com ODoE e KS e o método de busca ativa e as Figuras 5.12 (b) e 5.13 (b) mostram os resultados obtidos para busca ativa utilizando a técnica GPR com e sem critério de parada.

O melhor resultado da busca ativa foi obtido ao se utilizar a técnica GPR, porém o tempo de execução é maior do que das demais técnicas, dessa forma utilizou-se o critério de parada (discutido anteriormente) para reduzir o tempo da busca. Ao se analisar o critério de parada, implementado na técnica GRP, nota-se que esse foi eficiente, pois o valor ótimo foi encontrado para a maioria dos subconjuntos, conforme mostram as Figuras 5.14 (a) e 5.15 (a). Os histogramas, conforme Figuras 5.14 (b) e 5.15 (b), mostram o valor ótimo encontrado, quando aplicado o critério de parada; o eixo horizontal mostra os valores assumidos pela variável de interesse e o eixo vertical mostra a frequência de observações dos valores que pertencem a esse

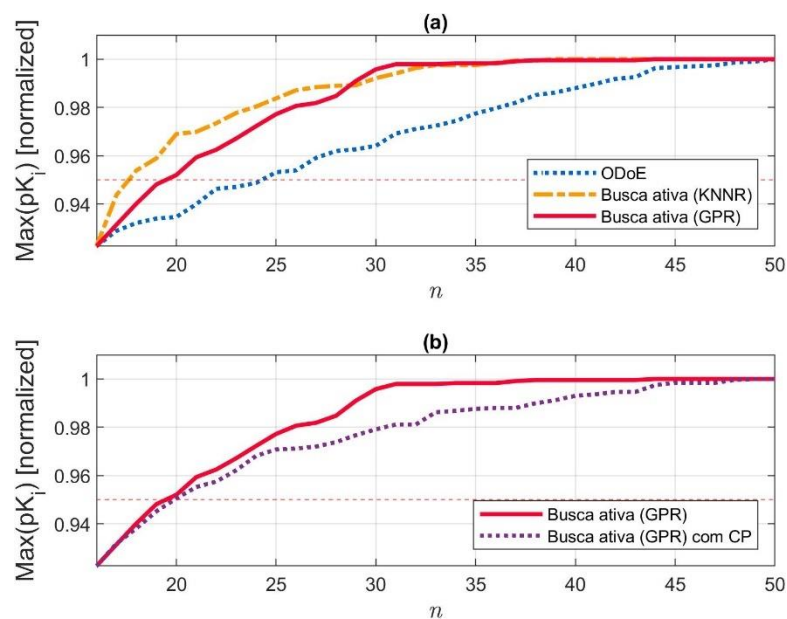
intervalo. Os histogramas, das Figuras 5.14 (c) e 5.15 (c), mostram a frequência do número de objetos selecionados (tamanho do conjunto i_{SEL}) no ponto de parada.

Figura 5.12. Resultado do método de busca ativa para a base de dados QSAR Medicamento, com a seleção dos objetos iniciais feito pelo KS.



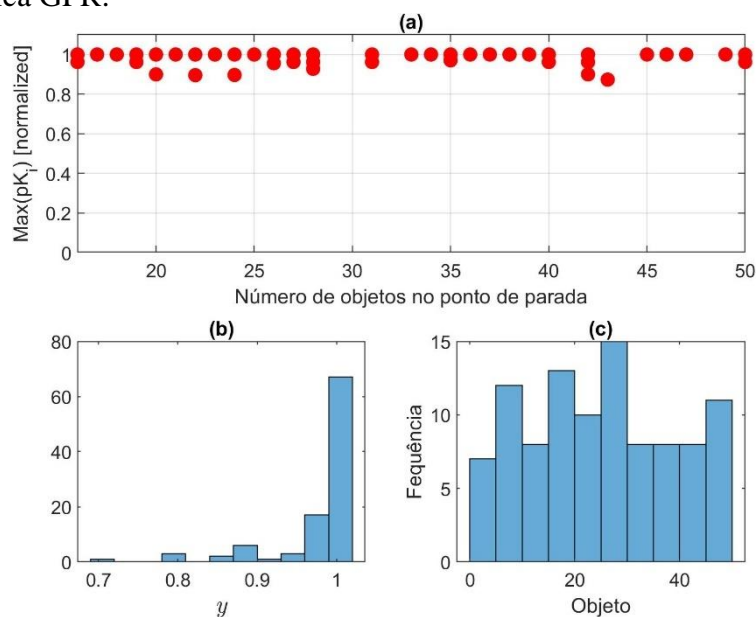
Fonte: A autora.

Figura 5.13. Resultado do método de busca ativa com a seleção dos objetos iniciais feito pelo ODoE.



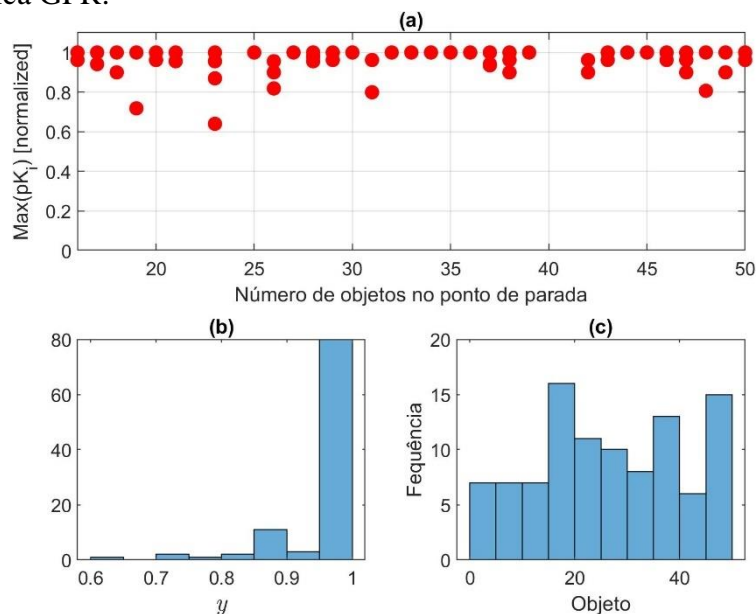
Fonte: A autora.

Figura 5.14. Ponto ótimo da base de dados QSAR Medicamentos utilizando seleção dos objetos iniciais pelo método KS e busca ativa com critério de parada para a técnica GPR.



Fonte: A autora.

Figura 5.15. Ponto ótimo da base de dados QSAR Medicamentos utilizando seleção dos objetos iniciais pelo método ODoE e busca ativa com critério de parada para a técnica GPR.



Fonte: A autora.

Além dessa análise, foi medido o tempo médio de execução dos algoritmos, conforme Tabela 5.2. Note que o critério de parada reduziu significativamente o tempo de execução do algoritmo GPR. O tempo de processamento foi reduzido em 96,7% utilizando ODoE como algoritmo de seleção inicial e busca ativa com critério de parada. De forma similar, o tempo de

processamento foi reduzido em 97.3% utilizando KS como algoritmo de seleção inicial e busca ativa com critério de parada. Essa redução do tempo não alterou a eficiência do método, pois os resultados são similares para a técnica GPR sem e com o critério de parada, conforme apresentado.

Tabela 5.2. Tempo médio de execução dos algoritmos para a base de dados QSAR Medicamento.

Método	Tempo médio de execução (segundos)
KS	0,022
KS e busca ativa <i>k</i> NNR	0,008
KS e busca ativa com GPR	2,280
KS e busca ativa com GPR e critério de parada	0,062
ODoE	0,111
ODoE e busca ativa <i>k</i> NNR	0,009
ODoE e busca ativa com GPR	2,341
ODoE e busca ativa com GPR e critério de parada	0,077

Fonte: A autora.

Desta forma, conforme resultados apresentados para este estudo, a busca ativa se apresenta como uma estratégia a ser considerada para o projeto experimental de um medicamento na fase inicial, na qual se deseja buscar a propriedade ótima utilizando poucos compostos.

5.4 QSAR Toxicidade

A avaliação do risco de produtos químicos para a saúde humana e o meio ambiente é uma medida importante para regulamentar seu uso seguro (BASANT; GUPTA, 2017; REN *et al.*, 2016). Nesse sentido, a toxicidade aquática de produtos químicos é uma das características importantes que têm um impacto direto/indireto sobre sua produção e uso (JIA *et al.*, 2018). O peixe é um dos principais representantes dos consumidores secundários (carnívoros) nas cadeias alimentares, por isso diversas espécies são utilizadas como bioindicadores. Algumas espécies são usadas para avaliar a toxicidade aquática de compostos orgânicos, dentre estas está o *Fathead minnow*. No Brasil, as espécies de peixe mais utilizadas são o *Danio rerio*, conhecido como peixe paulistinha ou peixe zebra e o *Pimephales promelas*, conhecido como *Fathead minnow*.

Os modelos QSAR têm sido utilizados para prever os possíveis efeitos toxicológicos de contaminantes (RODRIGUES, 2016), pois são alternativas ideais para experimentos por causa de sua maior eficiência e menor custo (WU; ZHANG; HU, 2016).

Neste estudo de caso, utilizou-se o QSAR de 908 moléculas orgânicas para prever a toxicidade aquática aguda em relação ao peixe *Pimepales promelas* (*Fathead Minnow*) extraídos da base disponível no Centro para Aprendizado de Máquina e Inteligência de Sistemas da Universidade da Califórnia (CASSOTTI *et al.*, 2015).

Para o QSAR toxicidade, foram calculados 06 (seis) descritores eletrônicos, conforme Quadro 5.2. Os valores dos descritores foram autoescalonados para garantir que as influências relativas de diferentes variáveis sejam independentes em suas unidades.

Quadro 5.2. Descritores eletrônicos e suas definições.

Descritor	Definição
MLOGP	Propriedades moleculares
CIC0	Índices de informação
GATS1i	Autocorrelações 2D
NdssC	Contagem de átomos
NdsCH	Contagem de átomos
SM1_Dz (Z)	Descritores baseados em matriz 2D

Fonte: A autora.

Em cada subconjunto foram selecionadas 100 moléculas (objetos) de forma aleatória. O método de pesquisa ativa proposto foi então aplicado a cada um destes subconjuntos, com o objetivo de encontrar o composto com o menor valor de concentração letal 50 (LC₅₀). Essa concentração causa 50% de mortalidade dos peixes, considerando uma exposição de 96 horas, é a mais frequentemente utilizada. Quanto maior o valor de LC₅₀, menor a toxicidade, pois é necessária uma concentração maior da substância tóxica para matar os peixes. Desse modo, quanto menor o [-LOG (mol/L)] de LC₅₀, menor a toxicidade. Neste estudo os valores [-LOG (mol/L)] de LC₅₀ variam entre 0,0529 e 9,6117.

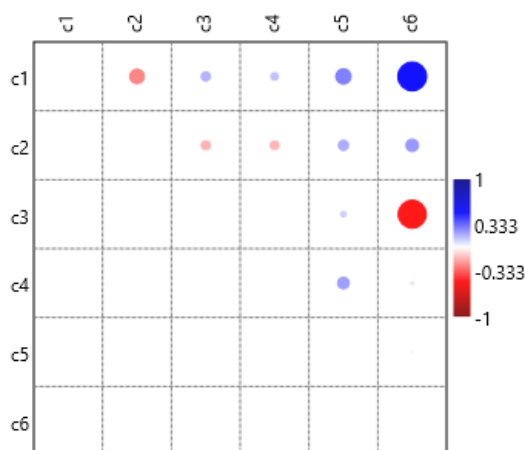
Também para esta base de dados, foi feita uma análise da correlação das componentes de **X**, conforme Tabela 5.3 e Figura 5.16. Nenhuma componente apresentou alta correlação com as demais. Desta forma, foram mantidas as 6 componentes. Foi utilizado o arranjo fatorial fracionado para cálculo do número de efeitos principais, similar à base de dados de QSAR medicamentos. Assim, se chegou ao valor de $n_o = 8$ (objetos iniciais).

Os resultados mostram que a busca ativa também convergiu rapidamente, conforme Figuras 5.17 e 5.18, respectivamente. As Figuras 5.17 (a) e 5.18 (a) mostram a comparação entre os métodos de busca ativa e as Figuras 5.17 (b) e 5.18 (b) mostram os resultados obtidos para busca ativa utilizando a técnica GPR com e sem critério de parada.

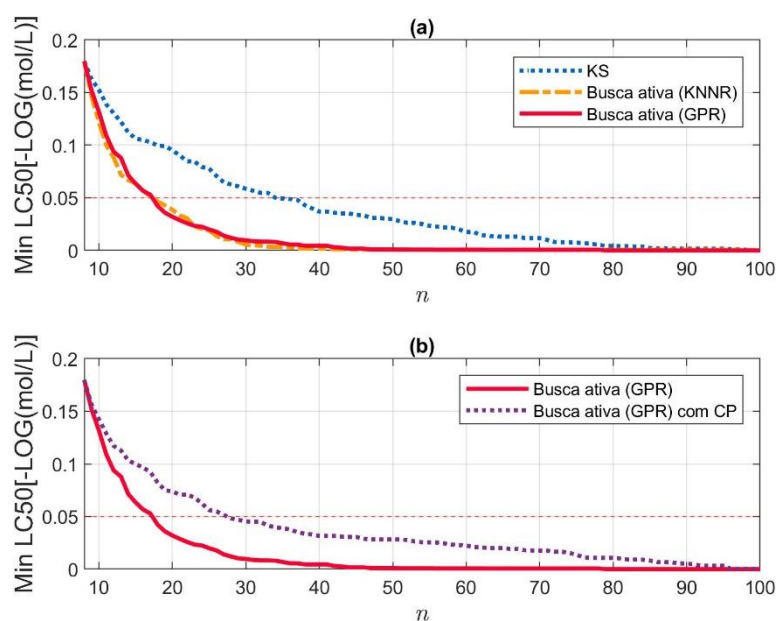
Tabela 5.3. Correlação entre as componentes do vetor x_i na base QSAR Toxicidade.

	c_1	c_2	c_3	c_4	c_5	c_6
c_1	1,00	-0,24	0,15	0,12	0,25	0,46
c_2	-0,24	1,00	-0,15	-0,14	0,16	0,20
c_3	0,15	-0,15	1,00	-0,01	0,09	-0,45
c_4	0,12	-0,14	-0,01	1,00	0,19	0,05
c_5	0,25	0,16	0,09	0,19	1,00	0,03
c_6	0,46	0,20	-0,45	0,05	0,03	1,00

Fonte: A autora.

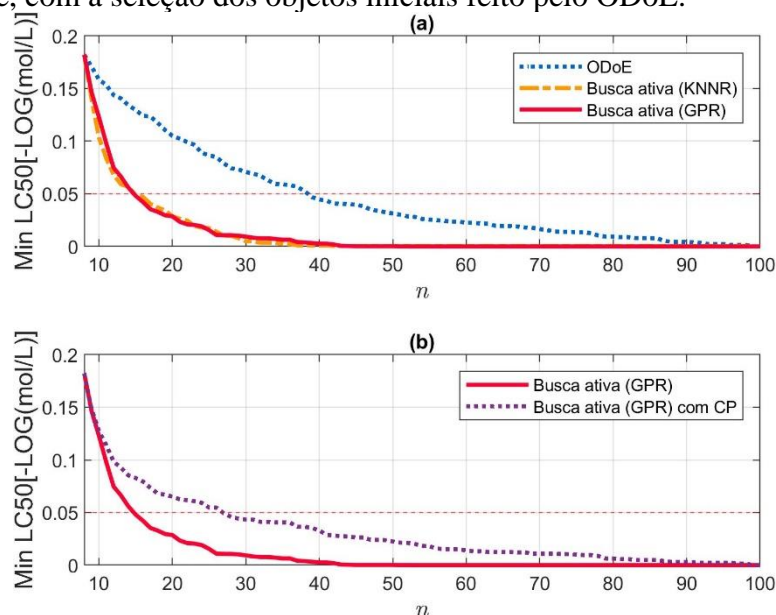
Figura 5.16. Correlação entre as componentes de X de QSAR Toxicidade.

Fonte: A autora.

Figura 5.17. Resultado do método de busca ativa para a base de dados QSAR Toxicidade, com a seleção dos objetos iniciais feito pelo KS.

Fonte: A autora.

Figura 5.18. Resultado do método de busca ativa para a base de dados QSAR Toxicidade, com a seleção dos objetos iniciais feito pelo ODoE.



Fonte: A autora.

Note que valores próximos ao ótimo (abaixo de 0,05) foram obtidos com menos de 20 objetos selecionados para a maioria dos conjuntos amostrais, o que mostra a eficiência do método proposto com objetos iniciais selecionamos por meio do KS e do ODoE

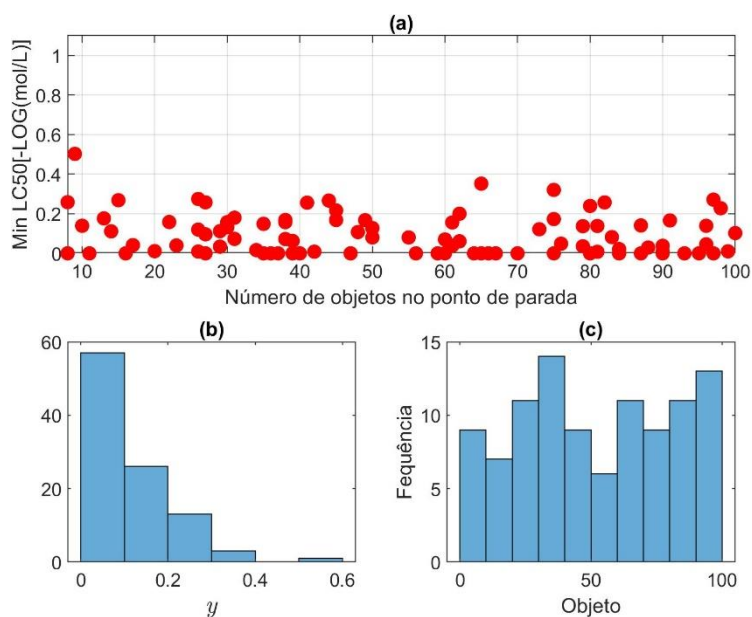
Ao se analisar o critério de parada, pode-se concluir que foi eficiente, pois o valor ótimo foi encontrado para a maioria dos subconjuntos, conforme Figuras 5.19 (a) e 5.20 (a).

O histograma à esquerda, conforme Figuras 5.19 (b) e 5.20 (b), mostram os valores ótimos encontrados para a concentração letal, quando aplicado o critério de parada (dentre os objetos selecionados).

O eixo horizontal mostra os valores assumidos pela variável de interesse e o eixo vertical mostra a frequência de observações dos valores que pertencem a esse intervalo. O histograma à direita, conforme Figuras 5.19 (c) e 5.20 (c), mostra a frequência do número de objetos selecionados no ponto de parada.

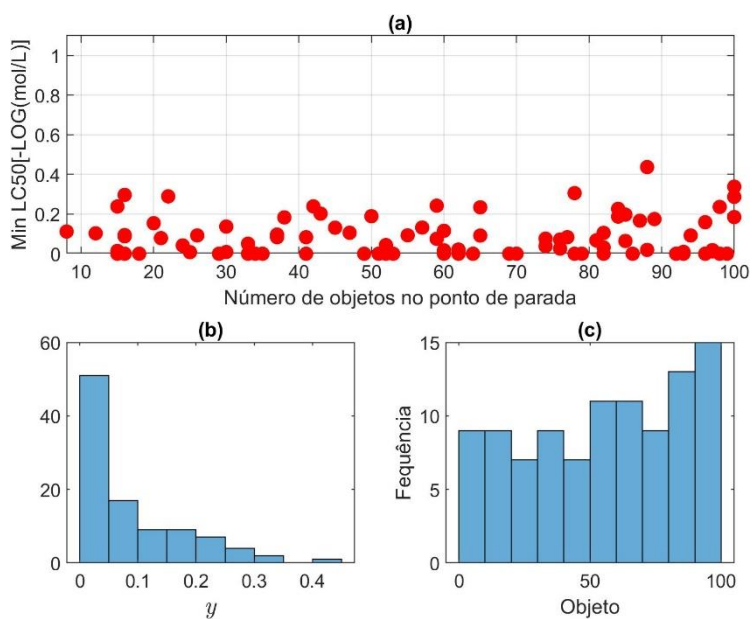
Ao se analisar o custo computacional, nota-se que este foi reduzido. A Tabela 6.4, similar ao estudo de caso anterior, apresenta o tempo médio de execução dos algoritmos. Ao se utilizar o critério de parada para a técnica GPR o tempo foi reduzido em 96,2% e 96,4%, sem alterar a eficiência do método.

Figura 5.19. Ponto ótimo da base de dados QSAR Toxicidade utilizando seleção dos objetos iniciais pelo método KS e busca ativa com critério de parada para a técnica GPR.



Fonte: A autora.

Figura 5.20. Ponto ótimo da base de dados QSAR Toxicidade utilizando seleção dos objetos iniciais pelo método ODoE e busca ativa com critério de parada para a técnica GPR.



Fonte: A autora.

Tabela 5.4. Tempo médio de execução dos algoritmos para a base de dados QSAR Toxicidade.

Método	Tempo médio de execução (segundos)
KS	0,009
KS e busca ativa KNNR	0,011
KS e busca ativa com GPR	4,248
KS e busca ativa com GPR e critério de parada	0,163
ODoE	0,291
ODoE e busca ativa KNNR	0,037
ODoE e busca ativa com GPR	4,112
ODoE e busca ativa com GPR e critério de parada	0,148

Fonte: A autora.

Além dos estudos referentes às bases de dados de QSAR, foi realizado um estudo de uma aplicação de busca ativa em microrredes elétricas. Os resultados são apresentados a seguir.

5.5 Aplicação em microrredes elétricas

Microrredes são sistemas de distribuição de energia elétrica contendo cargas e recursos de energia distribuída (RED) que podem operar conectadas à rede de energia principal ou ilhadas (EL KHAOUAT; BENHLIMA, 2020; FENG *et al.*, 2018; FERDOUS *et al.*, 2020). São exemplos de RED que podem ser utilizados na microrrede: geradores de fonte renovável (eólica, solar, hídrica, biomassa) e não renovável (baseadas em combustíveis fósseis, etc.); sistemas de armazenamento de energia (banco de baterias, volantes de inércia, sistemas de ar-comprimido, etc.) e veículos elétricos.

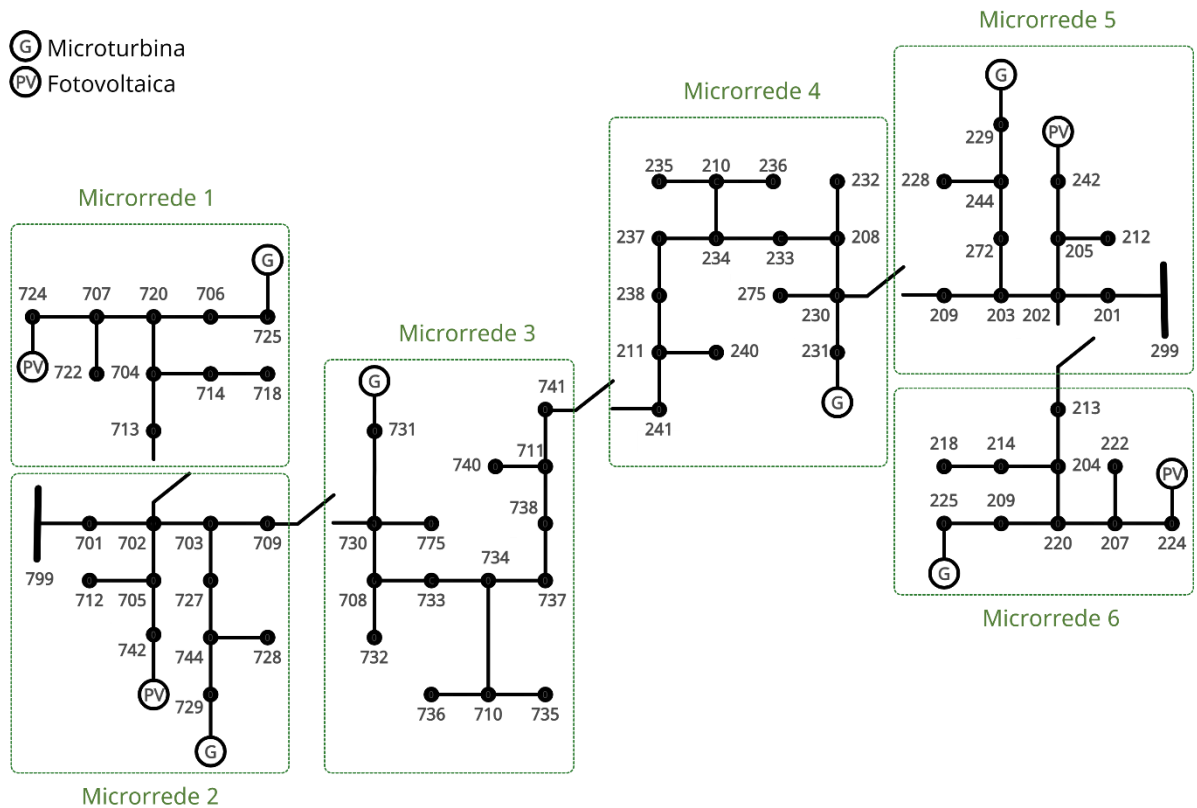
Para atender ao crescimento da implantação de microrredes, é essencial garantir que esses sistemas sejam altamente robustos em todas as condições de operação, ou seja, que permaneçam estáveis apesar das incertezas (PULCHERIO *et al.*, 2018). De acordo com os autores, essa capacidade de resistir a distúrbios e se adaptar a cenários de operação em mudança foi identificada como um aspecto-chave das futuras redes elétricas.

O conceito de agrupamento de microrredes pode ser aplicado à situação de acoplamento temporário de microrredes ilhadas, com o objetivo de aumentar a capacidade de serviço destas, durante o período de desconexão do sistema elétrico principal. A estratégia de agrupamento provê o atendimento às cargas prioritárias das microrredes, concomitantemente à manutenção dos níveis de tensão das barras elétricas e frequência do sistema dentro dos limites operativos.

A Figura 5.21 ilustra o sistema com seis microrredes no qual são realizados estudos de busca ativa. Cada microrrede corresponde à um sistema de distribuição radial padrão. Em operação normal, as microrredes operam conectadas ao sistema principal. Para as simulações

realizadas, assume-se que as microrredes estão isoladas do sistema elétrico principal, tal como ilustra a Figura.

Figura 5.21. Microrredes ilhadas do sistema principal.



Fonte: A autora.

O sistema de microrredes proposto possui fontes de energia despacháveis e não-despacháveis. As fontes de energia despacháveis possuem a sua potência de saída constantes, estando sujeitas às restrições técnicas, como limite de capacidade e disponibilidade de combustível (SILVA NETO, 2020). São exemplos de geração despacháveis: pequenas centrais hidrelétricas e termoelétricas. Neste trabalho, utiliza-se fontes de energia despacháveis nas barras 725, 729, 731, 225, 229 e 231. As unidades não despacháveis, ao contrário, tem geração de energia variável, já que a fonte de energia é dependente das condições climáticas, sendo intermitentes (SILVA NETO, 2020). Os principais exemplos de unidades não-despacháveis são as fontes de geração eólica e solar. Neste trabalho, utiliza-se apenas a geração fotovoltaica como fonte intermitente de energia nas barras 724, 742, 224 e 242.

5.5.1 Estudos preliminares para implementação da busca ativa

Aqui são apresentados estudos preliminares para definição dos procedimentos de implantação da metodologia de busca ativa em microrredes elétricas. Tal estudo auxilia no desenvolvimento de diretrizes de projetos futuros para implementação da metodologia de busca ativa para definição de agrupamentos de microrredes considerando-se diferentes aspectos operativos.

O cenário considerado para o estudo é a desconexão das microrredes, apresentadas na Figura 5.21, do sistema principal das 12h às 14h; assim, as microrredes ficam ilhadas por 2 horas. Neste período, a geração de energia de cada microrrede, bem como a demanda relativa à carga prioritária é apresentada na Tabela 5.5. Nota-se que as microrredes 3 e 4 possuem somente fontes de energia despacháveis. A potência base adotada é de 2500 VA; dessa forma, uma geração de, por exemplo, 2 *p.u.h* corresponde à 5kWh.

Tabela 5.5. Geração despachável e não-despachável das microrredes elétricas.

Microrrede	Geração despachável (p.u.h)	Geração não-despachável (p.u.h)	Demanda crítica (p.u.h)	Geração total (p.u.h)
M1	0,283	0,381	0,653	0,664
M2	1,638	0,457	2,068	0,095
M3	0,732	0	0,723	0,732
M4	0,732	0	0,723	0,732
M5	1,638	0,457	2,068	2,095
M6	0,283	0,381	0,653	0,664

Fonte: A autora.

Para esse estudo, a análise da melhor configuração considera que no momento da desconexão, a microrrede fornece energia apenas para as cargas classificadas como críticas, ou seja, cargas consideradas como não críticas são desligadas.

As seis microrredes ilhadas, apresentadas na Figura 5.21 (subseção anterior), podem ser combinadas em 32 agrupamentos distintos, conforme Tabela 5.6. Convém destacar que um agrupamento é composto de uma ou mais clusters e que uma *cluster* pode ser composta de uma ou mais microrredes. Nessa tabela, a primeira coluna é o número do objeto (Obj.), e as colunas relacionam-se às clusters 1-6. Uma cluster pode ter de uma microrrede ou mais, por exemplo, a cluster pode ter 6 microrredes agrupadas, como é o caso do Objeto 1. As microrredes 1-6 são representadas por M1, M2, M3, M4, M5 e M6, respectivamente.

Tabela 5.6. Agrupamentos de microrredes no sistema proposto.

<i>Clusters</i>						
Obj.	1	2	3	4	5	6
1	M1-M2-M3-M4-M5-M6					
2	M1-M2-M3-M4-M5	M6				
3	M1-M2-M3-M4	M5-M6				
4	M1-M2-M3-M4	M5	M6			
5	M1-M2-M3	M4-M5-M6				
6	M1-M2-M3	M4-M5	M6			
7	M1-M2-M3	M4	M5-M6			
8	M1-M2-M3	M4	M5	M6		
9	M1-M2	M3-M4-M5-M6				
10	M1-M2	M3-M4-M5	M6			
11	M1-M2	M3-M4	M5-M6			
12	M1-M2	M3-M4	M5	M6		
13	M1-M2	M3	M4-M5-M6			
14	M1-M2	M3	M4-M5	M6		
15	M1-M2	M3	M4	M5-M6		
16	M1-M2	M3	M4	M5	M6	
17	M1	M2-M3-M4-M5-M6				
18	M1	M2-M3-M4-M5	M6			
19	M1	M2-M3-M4	M5-M6			
20	M1	M2-M3-M4	M5	M6		
21	M1	M2-M3	M4-M5-M6			
22	M1	M2-M3	M4-M5	M6		
23	M1	M2-M3	M4	M5-M6		
24	M1	M2-M3	M4	M5	M6	
25	M1	M2	M3-M4-M5-M6			
26	M1	M2	M3-M4-M5	M6		
27	M1	M2	M3-M4	M5-M6		
28	M1	M2	M3-M4	M5	M6	
29	M1	M2	M3	M4-M5-M6		
30	M1	M2	M3	M4-M5	M6	
31	M1	M2	M3	M4	M5-M6	
32	M1	M2	M3	M4	M5	M6

Fonte: A autora.

Em continuidade, são detalhados como foram feitos os agrupamentos de microrredes para utilização da metodologia de busca ativa proposta:

1. Cada objeto, descrito da Tabela 5.6, corresponde a uma possibilidade de agrupamento.
2. De acordo com as conexões disponíveis, têm-se diferentes possibilidades de *clusters* e estas estão associadas às componentes do objeto.
3. Todos os objetos têm uma propriedade y relacionada que, neste caso, refere-se à perda total de energia elétrica do agrupamento.
4. O algoritmo de busca quer identificar o objeto que apresenta menor valor de y .

Sabe-se que cada objeto é uma possibilidade de agrupamento e é formado por componentes. Inicialmente, as componentes definidas foram a energia gerada e a energia demandada de cada *cluster*. Para maior entendimento, dois exemplos para a configuração de objetos são descritos a seguir:

1. Agrupamento 1 – que corresponde ao Objeto 1:

cluster 1 - (M1-M2-M2-M3-M4-M5-M6) / *cluster* 2 - vazia / ... / *cluster* 6 - vazia

Para este objeto as componentes são:

- c_1 : energia gerada pela *cluster* 1
- c_2 : energia demandada pela *cluster* 1
- c_3 : 0 (energia gerada pela *cluster* 2)
- c_4 : 0 (energia demandada pela *cluster* 2)
- c_5 : 0 (energia gerada pela *cluster* 3)
- c_6 : 0 (energia demandada pela *cluster* 3)
- c_7 : 0 (energia gerada pela *cluster* 4)
- c_8 : 0 (energia demandada pela *cluster* 4)
- c_9 : 0 (energia gerada pela *cluster* 5)
- c_{10} : 0 (energia demandada pela *cluster* 5)
- c_{11} : 0 (energia gerada pela *cluster* 6)
- c_{12} : 0 (energia demandada pela *cluster* 6)

2. Agrupamento 2 – que corresponde ao Objeto 2:

cluster 1- (M1-M2-M2-M3-M4-M5) / *cluster* 2 - M6 / *cluster* 3- vazia / ... / *cluster* 6 - vazia

Para este objeto as componentes são:

- c_1 : energia gerada pela *cluster* 1
- c_2 : energia demandada pela *cluster* 1
- c_3 : energia gerada pela *cluster* 2
- c_4 : energia demandada pela *cluster* 2
- c_5 : 0 (energia gerada pela *cluster* 3)

- c_6 : 0 (energia demandada pela *cluster* 3)
- c_7 : 0 (energia gerada pela *cluster* 4)
- c_8 : 0 (energia demandada pela *cluster* 4)
- c_9 : 0 (energia gerada pela *cluster* 5)
- c_{10} : 0 (energia demandada pela *cluster* 5)
- c_{11} : 0 (energia gerada pela *cluster* 6)
- c_{12} : 0 (energia demandada pela *cluster* 6)

Neste estudo, a busca ativa é utilizada para encontrar o agrupamento que tem a maior disponibilidade de energia para atender as cargas prioritárias quando o sistema estiver ilhado, ou seja, o agrupamento com a menor perda total de energia elétrica associada.

Na simulação (experimentação) destes sistemas, o cálculo de perda total de cada agrupamento requer a solução do problema de fluxo de potência de cada *cluster* de um dado agrupamento (a perda total corresponde à soma das perdas de cada *cluster*).

O fluxo de potência é a ferramenta mais utilizada por um engenheiro de uma empresa de energia elétrica, que trabalha na área de análise ou planejamento (MONTICELLI; GARCIA, 2011). O problema do fluxo de potência em cada rede de energia elétrica consiste na determinação do estado (tensões complexas das barras), na distribuição dos fluxos (potências ativas e reativas que fluem pelas linhas e transformadores) e de outras grandezas de interesse (MONTICELLI; GARCIA, 2011). O cálculo do fluxo de potência é realizado utilizando-se métodos computacionais desenvolvidos especificamente para a resolução de sistemas de equações e inequações algébricas que constituem o modelo estático da rede (tais como os métodos de Newton-Raphson e Gauss).

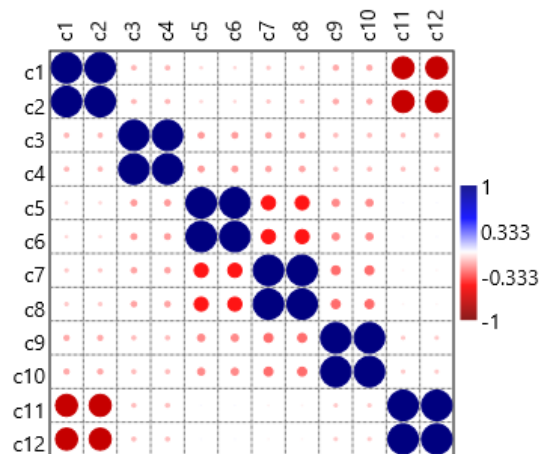
Similar aos estudos anteriores, foi feita uma análise da correlação das componentes de \mathbf{X} e foram retirados dos objetos as componentes $c_2, c_4, c_6, c_8, c_{10}, c_{11}$ e c_{12} (destacadas na Tabela 5.7) que possuem alta correlação, sobrando 5 componentes (c_1, c_3, c_5, c_7 e c_9) que correspondem à energia gerada de cada *cluster* de um agrupamento/objeto. Para ilustrar o problema, a Figura 6.22 mostra a correlação entres as componentes de forma gráfica.

Neste problema, se chegou ao valor de $n_o = 8$ (objetos iniciais), considerando o cálculo do número de experimentos iniciais, o mesmo utilizado para o fatorial fracionado. Nesta aplicação, cada objeto consiste em um agrupamento dentre os agrupamentos possíveis considerados para as microrredes.

Tabela 5.7. Correlação entre as componentes definidas para representas as microrredes.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
c_1	1,00	1,00	-0,30	-0,30	-0,39	-0,39	-0,50	-0,50	-0,27	-0,27	1,00	1,00
c_2	1,00	1,00	-0,30	-0,30	-0,39	-0,39	-0,50	-0,50	-0,27	-0,27	1,00	1,00
c_3	-0,30	-0,30	1,00	1,00	-0,47	-0,47	-0,24	-0,24	-0,18	-0,18	-0,30	-0,30
c_4	-0,30	-0,30	1,00	1,00	-0,47	-0,48	-0,24	-0,24	-0,18	-0,18	-0,30	-0,30
c_5	-0,39	-0,39	-0,47	-0,47	1,00	1,00	0,06	0,06	0,01	0,01	-0,39	-0,39
c_6	-0,39	-0,39	-0,47	-0,48	1,00	1,00	0,07	0,06	0,01	0,01	-0,39	-0,39
c_7	-0,50	-0,50	-0,24	-0,24	0,06	0,07	1,00	1,00	0,04	0,04	-0,50	-0,50
c_8	-0,50	-0,50	-0,24	-0,24	0,06	0,06	1,00	1,00	0,04	0,04	-0,50	-0,50
c_9	-0,27	-0,27	-0,18	-0,18	0,01	0,01	0,04	0,04	1,00	1,00	-0,27	-0,27
c_{10}	-0,27	-0,27	-0,18	-0,18	0,01	0,01	0,04	0,04	1,00	1,00	-0,27	-0,27
c_{11}	1,00	1,00	-0,30	-0,30	-0,39	-0,39	-0,50	-0,50	-0,27	-0,27	1,00	1,00
c_{12}	1,00	1,00	-0,30	-0,30	-0,39	-0,39	-0,50	-0,50	-0,27	-0,27	1,00	1,00

Fonte: A autora.

Figura 5.22. Correlação entre as componentes de X que representam as microrredes.

Fonte: A autora.

Após este estudo inicial, são mostrados os resultados obtidos com as simulações realizadas por meio de dois estudos, apresentados nas próximas seções.

5.5.2 Simulação com todos os agrupamentos possíveis para as microrredes

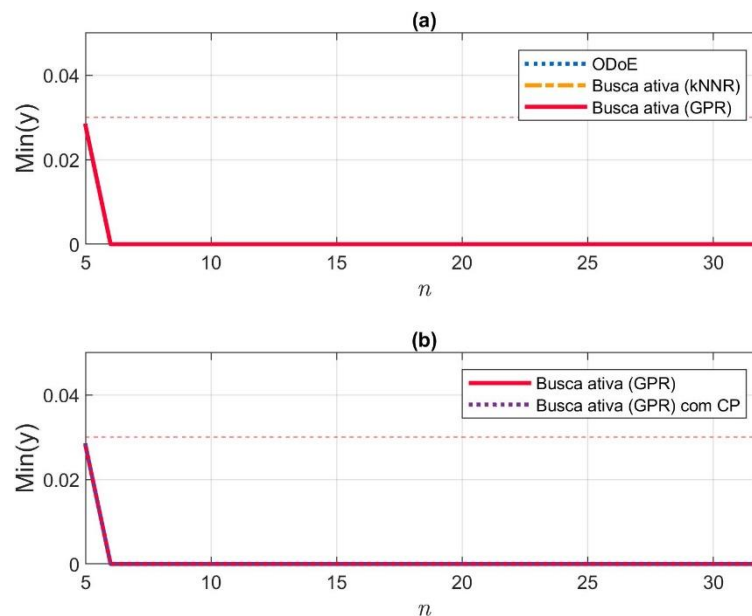
O primeiro estudo considerou os 32 possíveis agrupamentos de microrredes para utilizar a metodologia de busca ativa.

A Figura 5.23 mostra o resultado do método de busca ativa para a determinação dos agrupamentos de microrredes, com a seleção dos objetos iniciais feito pelo ODoE. O método de seleção inicial encontrou o melhor agrupamento na 6ª interação. Este objeto corresponde ao

32º objeto. Conseqüentemente, busca ativa também encontrou o valor ótimo, na seleção do 6º objeto, para ambas as técnicas empregadas (*k*NNR e GPR). Esse objeto corresponde a um agrupamento no qual não há interconexão alguma entre as microrredes ilhadas. Esta configuração apresenta perdas totais de energia de 0.0952 *p.u.h*. Assim, o método de busca ativa convergiu rapidamente.

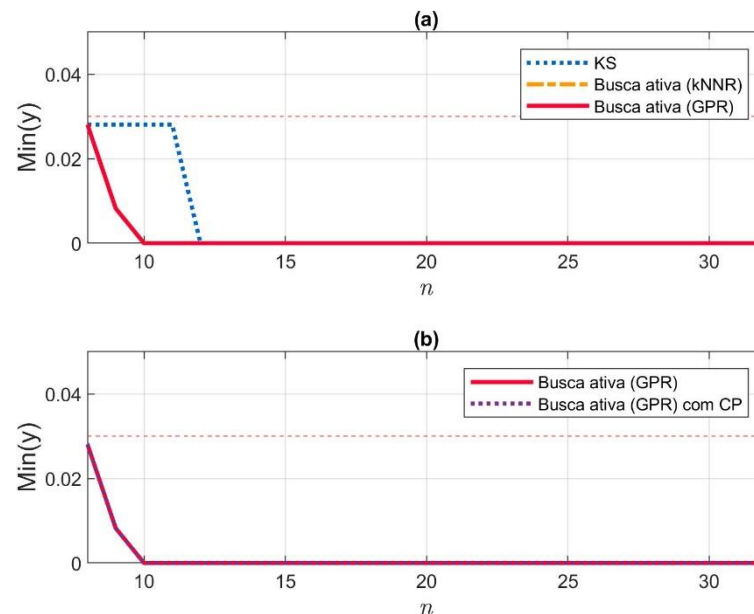
O resultado do método de busca ativa para a determinação dos agrupamentos de microrredes, com a seleção dos objetos iniciais feito pelo KS, também encontra o melhor agrupamento rapidamente, de acordo com a Figura 5.24. O agrupamento com menor perda de energia foi encontrado pela busca ativa na avaliação do 10º objeto para as técnicas *k*NNR e GPR. Se utilizarmos somente a busca pelo algoritmo KS, este encontra o melhor agrupamento na 12ª interação. A Figura 5.24 (a) mostra a comparação entre o método KS e a busca ativa e a Figura 5.24 (b) faz comparação dos resultados obtidos para busca ativa utilizando a técnica GPR com e sem critério de parada.

Figura 5.23. Resultado do método de busca ativa considerando 32 possíveis agrupamentos de microrredes, com a seleção dos objetos iniciais feito pelo ODoE.



Fonte: A autora.

Figura 5.24. Resultado do método de busca ativa considerando 32 possíveis agrupamentos de microrredes, com a seleção dos objetos iniciais feito pelo KS.



Fonte: A autora.

Para um segundo estudo, foi retirado o objeto 32, que corresponde ao agrupamento com menor perda de energia. Os resultados são apresentados na próxima seção.

5.5.3 Simulação com 31 agrupamentos para as microrredes

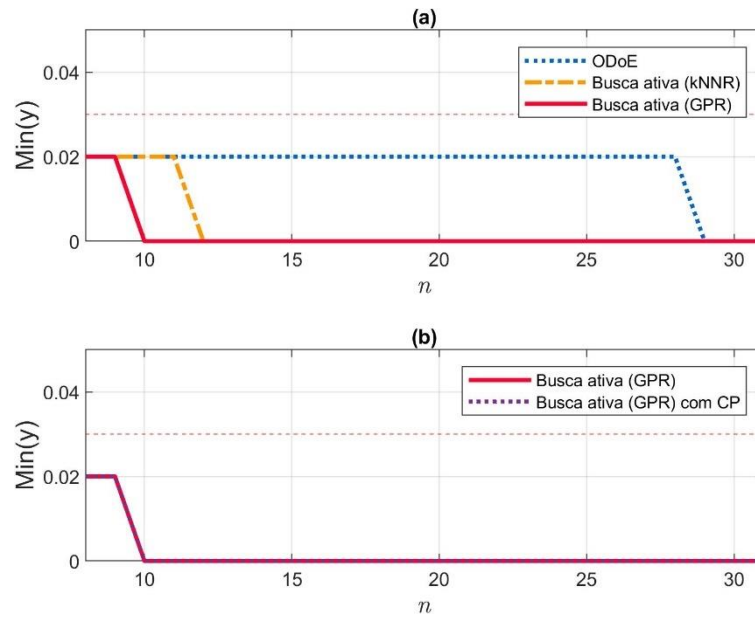
O segundo estudo de simulação considera 31 possíveis agrupamentos entre as microrredes. Foi retirado dessa simulação o objeto 32. Esse Objeto corresponde ao agrupamento que melhor atende à propriedade de interesse (menor perda de energia) no qual todas as microrredes apresentam-se isoladas. Admite-se, dessa forma, a situação prática na qual um particular agrupamento não possa ser implementado.

A Figura 6.25 mostra o resultado do método de busca ativa, com a seleção dos objetos iniciais feito pelo ODoE. Neste contexto, a busca ativa com a técnica GPR sem e com critérios de parada apresentaram o melhor resultado, pois na seleção do 10º objeto encontraram o melhor agrupamento para as microrredes ilhadas. A busca ativa com a técnica $kNNR$ encontrou o melhor agrupamento no 12º objeto selecionado e o ODoE no 29º objeto selecionado. A Figura 5.25 (a) mostra comparação entre o método KS e a busca ativa e a Figura 5.25 (b) faz comparação dos resultados obtidos para busca ativa utilizando a técnica GPR sem e com critério de parada.

Ao repetir a simulação, o melhor agrupamento de microrredes foi encontrado na avaliação do 9º objeto para todas as técnicas de busca ativa propostas nesta tese e o método KS

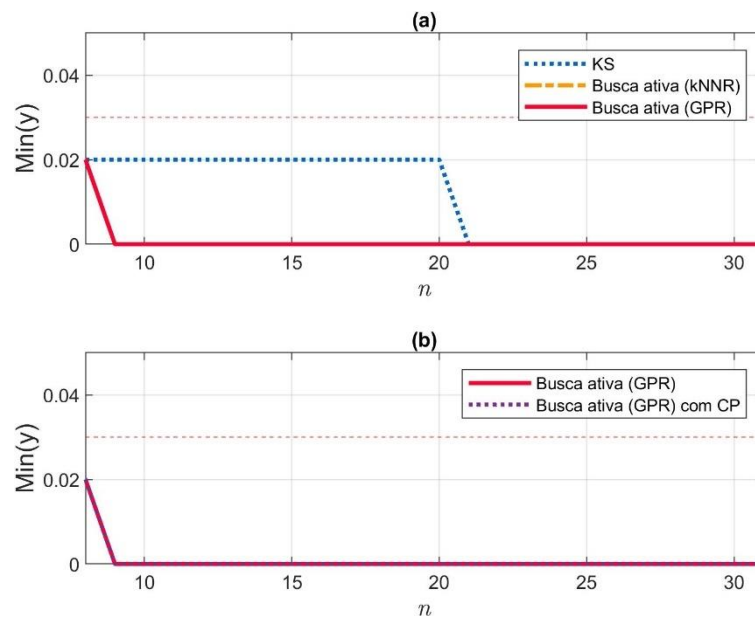
encontrou a melhor configuração na avaliação do 21º objeto, conforme Figura 5.26. A Figura 5.26 (a) mostra comparação entre o método KS e a busca ativa e a Figura 5.26 (b) faz comparação dos resultados obtidos para busca ativa utilizando a técnica GPR com e sem critério de parada.

Figura 5.25. Resultado do método de busca ativa considerando 31 possíveis agrupamentos de microrredes, com a seleção dos objetos iniciais feito pelo ODoE.



Fonte: A autora.

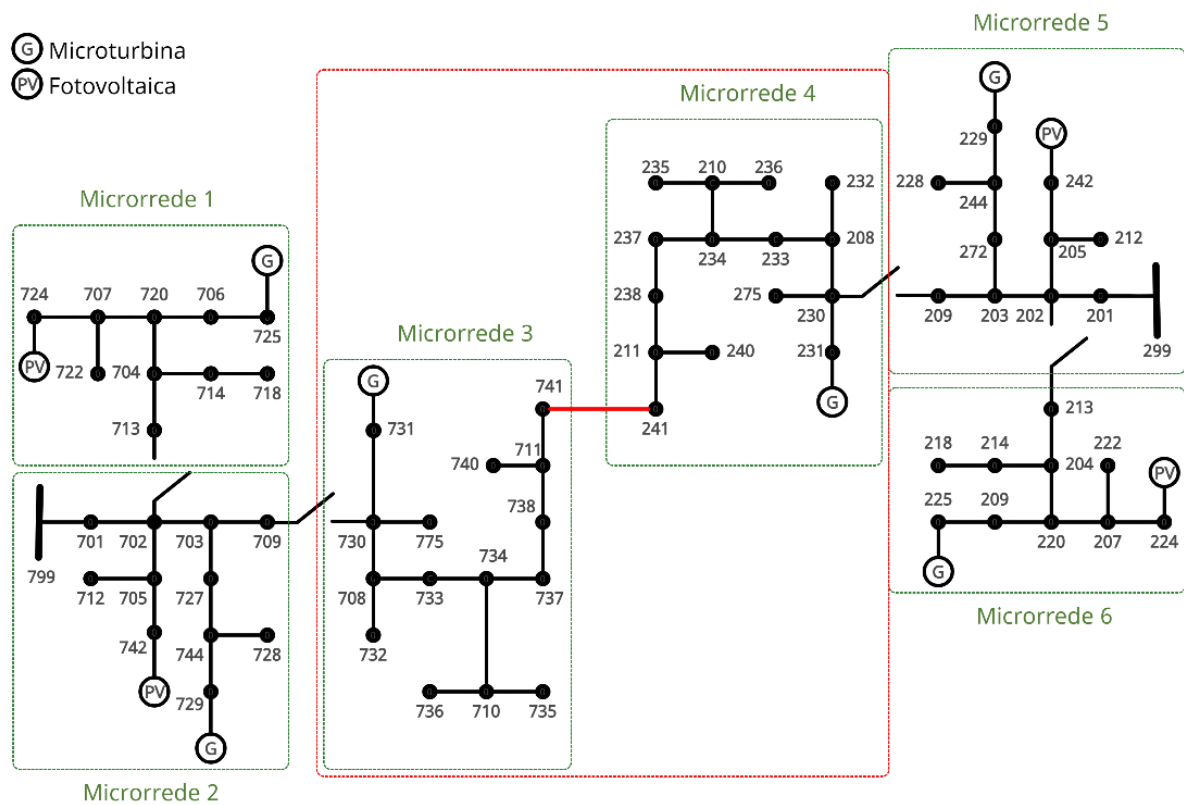
Figura 5.26. Resultado do método de busca ativa considerando 31 possíveis agrupamentos de microrredes, com a seleção dos objetos iniciais feito pelo KS.



Fonte: A autora.

O agrupamento de microrrede que melhor atende à propriedade de interesse corresponde ao objeto 28, no qual as Microrredes 1, 2, 5 e 6 encontram-se isoladas e as Microrredes 3 e 4 encontram-se agrupadas, de acordo com a Figura 5.27. Convém destacar que essas últimas são as microrredes que possuem somente geração não-despachável.

Figura 5.27. Agrupamento de microrredes após a aplicação do algoritmo de busca.



Fonte: A autora

Após a realização destes estudos, os resultados apontam que a busca ativa se apresenta potencialmente interessante para aplicação no problema de definição de agrupamento de microrredes ilhadas.

5.6 Análises finais

Nesta seção são analisados os resultados obtidos para a metodologia de busca ativa proposta. O desempenho da busca ativa com k NNR e GPR foi avaliado e comparado com os métodos tradicionais KS e ODoE de busca, para três bases de dados. Convém destacar que esses últimos, também foram utilizados na seleção dos n_0 objetos, na metodologia de busca ativa. Primeiro, nota-se que, em todos os estudos realizados, a busca ativa tem um desempenho melhor se comparado aos algoritmos tradicionais de busca ODoE e KS.

Conforme pode-se observar na Tabela 5.8, a busca ativa minimizou substancialmente o número de experimentos. Na tabela, são mostradas as bases de dados utilizadas nas simulações, o número de objetos, o número inicial de objetos, os métodos de busca utilizados, a quantidade de objetos selecionados para encontrar a propriedade de interesse com valor ótimo e, por fim, quanto o algoritmo de busca minimizou o número de experimentos.

Tabela 5.8. Número de experimentos após a aplicação do algoritmo de busca

Base de dados	Objetos	n_0	Método	Experimentos	
				Ótimo	Redução
QSAR Medicamento	50	16	ODoE	24*	52,0%
			Busca ativa com ODoE e k NNR	18*	64,0%
			Busca ativa com ODoE e GPR	19*	62,0%
			KS	18*	64,0%
			Busca ativa com KS e k NNR	17*	66,0%
			Busca ativa com KS e GPR	17*	66,0%
QSAR Toxicidade	100	8	ODoE	39*	61,0%
			Busca ativa com ODoE e k NNR	15*	85,0%
			Busca ativa com ODoE e GPR	15*	85,0%
			KS	35*	65,0%
			Busca ativa com KS e k NNR	18*	82,0%
			Busca ativa com KS e GPR	18*	82,0%
Microrredes	32	8	ODoE	6	81,3%
			Busca ativa com ODoE e k NNR	6	81,3%
			Busca ativa com ODoE e GPR	6	81,3%
			KS	12	62,5%
			Busca ativa com KS e k NNR	10	68,8%
	31	8	Busca ativa com KS e GPR	10	68,8%
			ODoE	29	6,5%
			Busca ativa com ODoE e k NNR	12	61,3%
			Busca ativa com ODoE e GPR	10	67,7%
			KS	21	32,3%
Busca ativa com KS e k NNR	09	71,0%			
Busca ativa com KS e GPR	09	71,0%			

*Média dos objetos com valores ótimos.

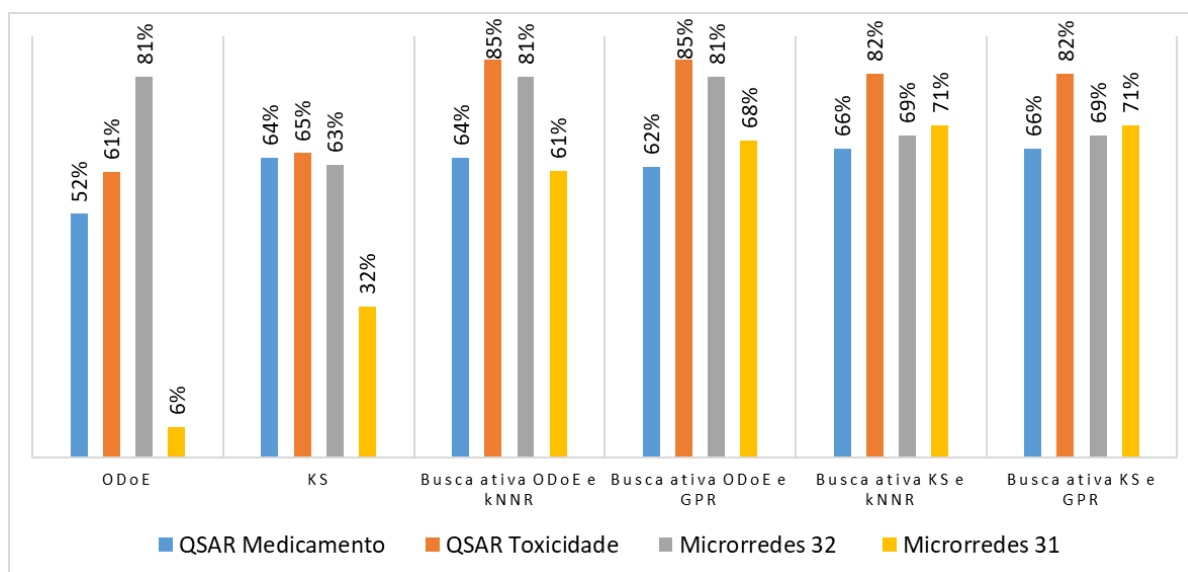
Fonte: A autora.

Convém ressaltar que para as bases de dados de QSAR Medicamentos e QSAR Toxidade foram apresentados os valores médios do número de objetos selecionados, pois foram criados 100 subconjuntos contendo 50 e 100 amostras respectivamente, conforme explicitado na metodologia do trabalho.

Nota-se que, para todas as bases de dados exploradas, houve uma redução significativa no número de experimentos para encontrar a propriedade de interesse. O gráfico da Figura 5.28 mostra a porcentagem da redução do número de experimentos ou simulações.

Para a base QSAR Medicamentos, buscou-se encontrar o composto com o maior valor de pK_i em cada subconjunto. Nesse caso, o número de experimentos necessários reduziu para menos de 20 objetos, o que corresponde à uma redução de mais de 60,0% utilizando os métodos de busca ativa.

Figura 5.28. Redução do número de experimentos ou simulações com a implementação do método de busca ativa.



Fonte: A autora.

Para a base QSAR Toxicidade, o objetivo foi encontrar o composto com o menor valor de concentração letal 50 (LC_{50}). Houve uma redução do número de experimentos de, pelo menos, 82,0% com o uso da busca ativa.

A taxa de redução do número de simulações para microrredes foi de pelo menos 81,0% e 61,0% para 32 e 31 configurações possíveis, respectivamente. Neste caso, há uma redução do cálculo para encontrar as perdas associadas ao agrupamento. No entanto, como trabalhos futuros, esses cálculos poderão ser implementados no algoritmo proposto e desta forma, realizar essa busca em tempo real.

Nota-se que o valor ótimo da propriedade foi encontrado após a escolha de um número reduzido de objetos e desta forma foi possível reduzir o trabalho experimental ou o número de simulações.

Ao resgatarmos o problema de pesquisa desta tese, pode-se dizer que foi possível desenvolver uma abordagem estruturada para busca de valores ótimos utilizando inteligência

artificial, baseada no paradigma de aprendizagem ativa de máquina e com isso minimizar o número de experimentos para um dado problema.

A regressão por k -vizinhos-mais-próximos é um tipo de algoritmos de aprendizagem de máquina robusto, simples de implementar e computacionalmente não custoso (CHU *et al.*, 2021; DURBIN *et al.*, 2021). O k NNR prevê o valor da propriedade de interesse a partir dos valores dos k objetos mais próximos, pode combinar o valor estimado (\hat{y}) para uma região limitada e, com isso, evitar o problema de sobre-ajuste (*overfitting*), possibilitando que o modelo tenha a capacidade de generalização (DURBIN *et al.*, 2021).

A regressão por processos Gaussianos é uma abordagem de otimização Bayesiana, na qual um modelo não linear e não paramétrico, com incerteza de predição quantificável, é ajustado para medições existentes (LEE, 2019). Para encontrar o próximo valor usa-se análise como base a dependência de estados, concentrando esforços de busca onde a resposta já foi observada como favorável, ou seja, diminui o espaço de busca por inferir predições a partir de solução promissoras. A técnica GPR é, portanto, adequada para problemas não lineares. Uma desvantagem desta abordagem é que os modelos podem ser difíceis de interpretar, uma vez que são empíricos e não paramétricos (LEE, 2019). Além disso, a configuração dos hiperparâmetros influencia o desempenho do algoritmo.

Após a apresentação e discussão dos resultados, são apresentadas as conclusões e sugestões para trabalhos futuros no capítulo a seguir.

6 Conclusões

A principal contribuição deste trabalho é o desenvolvimento de métodos de busca ativa para encontrar objetos com propriedades ótimas referentes à propriedade y , com o intuito de reduzir a carga de trabalho experimental. Outra contribuição foi a utilização da técnica de regressão por processos Gaussianos e da definição de um critério de parada para a mesma que reduziu substancialmente o custo computacional.

Com o advento dos métodos numéricos e facilidades computacionais, tornou-se possível projetar experimentos por meio de simulações baseadas em computador. Selecionar um bom conjunto de objetos (também chamados de observações, pontos de treinamento ou amostras) tornou-se uma questão fundamental em simulações baseadas em computador, com o objetivo predominante de maximizar a quantidade de informações obtidas a partir de um número limitado de objetos (YONDO; ANDRÉS; VALERO, 2018). Desta forma, o uso do ODoE e do KS se mostrou uma boa forma de seleção inicial para favorecer a busca ativa.

Na técnica k NNR, o número de k , empregados na estimativa da propriedade y , é um parâmetro de projeto que precisa ser escolhido pelo analista (experimentador/pesquisador). No entanto, os testes realizados em trabalho anterior (MATTA *et al.*, 2016) sugerem que essa escolha não é um fator crítico para a aplicação do método proposto. Na verdade, bons resultados também foram obtidos usando uma versão mais simples dessa técnica, na qual todos os objetos já selecionados são empregados na estimativa de y .

A técnica GPR apresentou bons resultados para otimizar uma função, porém não se mostrou tão eficiente em relação ao tempo de processamento, devido a isso, foi implementado um critério de parada por meio do cálculo dos intervalos de confiança para os valores estimados \hat{y} .

Foram apresentados exemplos com dados simulados, bem como conjuntos de dados reais de QSAR Medicamentos e Toxicidade. Foram realizados também estudos e simulações para configuração de microrredes elétricas ilhadas. Em todos estes casos, a busca ativa produziu resultados com valores ótimos ou próximos ao ótimo, ou seja, resultados próximos ao valor máximo de y . Assim, a busca ativa atingiu o objetivo de reduzir o número de experimentos para encontrar o objeto ideal que possuir o valor ótimo para a propriedade y desejada.

Além disso, pode-se concluir que os objetivos da tese foram atingidos, pois foi possível (1) utilizar delineamento de experimentos e do algoritmo Kenard-Stone para seleção inicial dos objetos que apresentam as melhores propriedades de interesse, (2) desenvolver uma abordagem de busca ativa destinada a buscar objetos com propriedades ótimas, (3) implementar

computacionalmente a técnica em um conjunto de dados simulados e em problemas reais e (4) avaliar os benefícios da metodologia de busca ativa proposta.

Uma limitação apresentada pelo trabalho foi que só pode ser predito uma propriedade de interesse. Outra limitação foi a configuração do hiperparâmetros da regressão por processos Gaussianos devido às várias possibilidades que esses oferecem.

Para trabalhos futuros sugere-se o a implementação de busca ativa utilizando outras técnicas de regressão de dados tais como regressão adaptativa multivariada *splines* (*Multivariate Adaptive Regression Spline* - MARS) e redes neurais artificiais. Além disso, um melhor desempenho da busca ativa com GPR pode ser esperado se alguns hiperparâmetros da técnica GRP forem otimizados.

Referências

- ALIZADEH, R.; ALLEN, J. K.; MISTREE, F. Managing computational complexity using surrogate models: a critical review. **Research in Engineering Design**, v. 31, n. 3, p. 275–298, 2020.
- ASFAW, B. A. et al. Optimization of compound-specific chlorine stable isotope analysis of chloroform using the Taguchi design of experiments. **Rapid Communications in Mass Spectrometry**, v. 34, n. 23, 2020.
- BALLABIO, C. et al. Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. **Geoderma**, v. 355, 2019.
- BARROS NETO, B. DE; SCARMINIO, I. S.; BRUNS, R. E. **Como fazer experimentos: pesquisa e desenvolvimento na ciência e na indústria**. São Paulo: Atmed, 2010.
- BASANT, N.; GUPTA, S. QSAR modeling for predicting mutagenic toxicity of diverse chemicals for regulatory purposes. **Environmental Science and Pollution Research**, v. 24, n. 16, p. 14430–14444, 2017.
- BROCHU, E.; CORA, V. M.; DE FREITAS, N. **A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning**. 2010.
- CASSOTTI, M. et al. A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (*Pimephales promelas*). **SAR and QSAR in environmental research**, v. 26, n. 3, p. 217–243, 2015.
- CHABANET, S.; BRIL EL-HAOUZI, H.; THOMAS, P. Coupling digital simulation and machine learning metamodel through an active learning approach in Industry 4.0 context. **Computers in Industry**, v. 133, 2021.
- CHAIB, A. E. et al. Optimal power flow with emission and non-smooth cost functions using backtracking search optimization algorithm. **International Journal of Electrical Power & Energy Systems**, v. 81, p. 64–77, 2016.
- CHEN, R.; IMANI, M.; IMANI, F. Joint active search and neuromorphic computing for efficient data exploitation and monitoring in additive manufacturing. **Journal of manufacturing processes**, v. 71, p. 743–752, 2021.
- CHU, J. et al. A novel method overcoming overfitting of artificial neural network for accurate prediction: Application on thermophysical property of natural gas. **Case Studies in Thermal Engineering**, v. 28, p. 101406, 2021.
- CORDEIRO, Margareth Moreira. **Fator de Bayes a posteriori para comparar os coeficientes de modelos de regressão**. 1993. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 1993. doi:10.11606/D.55.2018.tde-27082018-101228.
- DAL BIANCO, G.; GONCALVES, M. A.; DUARTE, D. BLOSS: Effective meta-blocking with almost no effort. **Information Systems**, v. 75, p. 75–89, 2018.
- DENG, Z. W. et al. Data-driven state of charge estimation for lithium-ion battery packs based on Gaussian process regression. **Energy**, v. 205, 2020.
- DEWAR, M. J. S. et al. AM1: A new general purpose quantum mechanical molecular model 1. **Journal of the American Chemical Society**, v. 107, n. 13, p. 3902–3909, 1985.

- DIMASI, J. A.; GRABOWSKI, H. G.; HANSEN, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. **Journal of Health Economics**, v. 47, p. 20–33, 2016.
- DING, Y. C. et al. Retrieving quantum information with active learning. **Physical Review Letters**, v. 124, n. 14, 2020.
- DOKE, S. K.; DHAWALE, S. C. Alternatives to animal testing: a review. **Saudi Pharmaceutical Journal**, v. 23, n. 3, p. 223–229, 2015.
- DURBIN, M. et al. K-nearest neighbors regression for the discrimination of gamma rays and neutrons in organic scintillators. **Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment**, v. 987, p. 164826, 2021.
- EL KHAOUAT, A.; BENHLIMA, L. A systematic literature review on prediction models in microgrids. **Journal of Theoretical and Applied Information Technology**, v. 98, n. 15, p. 3011–3029, 2020.
- FENG, W. et al. A review of microgrid development in the United States – A decade of progress on policies, demonstrations, controls, and software tools. **Applied Energy**, v. 228, p. 1656–1668, 2018.
- FERDOUS, S. M. et al. Dynamic frequency and overload management in autonomous coupled microgrids for self-healing and resiliency improvement. **IEEE access**, v. 8, p. 116796–116811, 2020.
- FINNEY, D. J. The fractional replication of factorial arrangements. **Annals of Eugenics**, v. 12, n. 4, p. 291–301, 1945.
- FISHER, R. A. The arrangement of field experiments. In: KOTZ, S.; JOHNSON, N. L. (Eds.). **Breakthroughs in statistics: methodology and distribution**. New York, NY: Springer New York, 1992. p. 82–91.
- FLETCHER, R. **Practical methods of optimization**. [s.l.] John Wiley & Sons, 1987.
- GARNETT, R. et al. Bayesian optimal active search and surveying. **arXiv.org**, 2012.
- GARNETT, R. et al. Introducing the “active search” method for iterative virtual screening. **Journal of Computer-Aided Molecular Design**, v. 29, n. 4, p. 305–314, 2015.
- GHESU, F. C. et al. Towards intelligent robust detection of anatomical structures in incomplete volumetric data. **Medical Image Analysis**, v. 48, p. 203–213, 2018.
- GUPTA, M.; SHARMA, R.; KUMAR, A. Docking techniques in pharmacology: How much promising? **Computational Biology and Chemistry**, v. 76, p. 210–217, 2018.
- GUVENIR, H. A.; UYSAL, I. Regression on feature projections. **Knowledge-Based Systems**, v. 13, n. 4, p. 207–214, 2000.
- HÄSE, F.; ROCH, L. M.; ASPURU-GUZIŁ, A. Next-generation experimentation with self-driving laboratories. **Trends in Chemistry**, v. 1, n. 3, p. 282–291, 2019.
- HERFURTH, H. **Gaussian process regression in computational finance**. [s.l.] Uppsala University, Applied Mathematics and Statistics, 2020.
- JIA, Q. et al. QSAR model for predicting the toxicity of organic compounds to fathead minnow. **Environmental Science and Pollution Research**, v. 25, n. 35, p. 35420–35428, 2018.
- JIANG, S. et al. Efficient nonmyopic active search with applications in drug and materials discovery. **arXiv.org**, 2018.

- JIANG, S. L.; MOSELEY, B.; GARNETT, R. Cost effective active search. **Advances in Neural Information Processing Systems 32 (Nips 2019)**, v. 32, 2019.
- KIANI, J. et al. Application of pool-based active learning in reducing the number of required response history analyses. **Computers & Structures**, v. 241, 2020.
- KIRKEY, A. et al. Optimization of the bulk heterojunction of all-small-molecule organic photovoltaics using design of experiment and machine learning approaches. **ACS Applied Materials & Interfaces**, v. 12, n. 49, p. 54596–54607, 2020.
- KONG, D. D.; CHEN, Y. J.; LI, N. Gaussian process regression for tool wear prediction. **Mechanical Systems and Signal Processing**, v. 104, p. 556–574, 2018.
- KOTHARI, R.; JAIN, V. Learning from labeled and unlabeled data using a minimal number of queries. **Ieee Transactions on Neural Networks**, v. 14, n. 6, p. 1496–1505, 2003.
- KUNDU, I.; PAUL, G.; BANERJEE, R. A machine learning approach towards the prediction of proteinligand binding affinity based on fundamental molecular properties. **RSC Advances**, v. 8, n. 22, p. 12127–12137, 2018.
- LEE, R. Statistical design of experiments for screening and optimization. **Chemie ingenieur technik**, v. 91, n. 3, p. 191–200, 2019.
- LEON, F.; CURTEANU, S. Large margin nearest neighbour regression using different optimization techniques. **Journal of Intelligent & Fuzzy Systems**, v. 32, n. 2, p. 1321–1332, 2017.
- LI, H.; DEL CASTILLO, E.; RUNGER, G. On active learning methods for manifold data. **Test**, v. 29, n. 1, p. 1–33, 2020.
- MARTÍNEZ-ESPARZA, J. et al. New 1-aryl-3-(4-arylpiperazin-1-yl)propane derivatives, with dual action at 5-HT 1A serotonin receptors and serotonin transporter, as a new class of antidepressants. **Journal of Medicinal Chemistry**, v. 44, n. 3, p. 418–428, 2001.
- MATTA, C. E. et al. An active search method for finding objects with near-optimal property values within a given set. **Journal of the Brazilian Chemical Society**, v. 27, n. 7, p. 1177–1187, 2016.
- MONTGOMERY, D. C. **Design and analysis of experiments**. 8. ed. Danvers, MA: John Wiley & Sons, 2013.
- MONTICELLI, A.; GARCIA, A. **Introdução a sistemas de energia elétrica**. Campinas: UNICAMP, 2011.
- MORAIS, C. L. M. et al. Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach. **Bioinformatics (Oxford, England)**, v. 35, n. 24, p. 5257–5263, 2019.
- NATH, A.; DE, P.; ROY, K. QSAR modelling of inhalation toxicity of diverse volatile organic molecules using no observed adverse effect concentration (NOAEC) as the endpoint. **Chemosphere**, v. 287, 2022.
- OGLIC, D. et al. Active search for computer-aided drug design. **Molecular Informatics**, v. 37, n. 1, p. n/a–n/a, 2018.
- ONAWOLE, A. T. et al. Structure based virtual screening of the Ebola virus trimeric glycoprotein using consensus scoring. **Computational Biology and Chemistry**, v. 72, p. 170–180, 2018.
- PETETTA, F.; CICCOCIOPPO, R. Public perception of laboratory animal testing: Historical,

- philosophical, and ethical view. **Addiction Biology**, v. n/a, n. n/a, p. e12991, 2020.
- PULCHERIO, M. et al. Robust microgrid clustering in a distribution system with inverter-based DERs. **IEEE transactions on industry applications**, v. 54, n. 5, p. 5152–5162, 2018.
- RAMIREZ-LOPEZ, L. et al. Sampling optimal calibration sets in soil infrared spectroscopy. **GEODERMA**, v. 226, p. 140–150, 2014.
- RASMUSSEN, C. E.; NICKISCH, H. Gaussian processes for machine learning (GPML) toolbox. **Journal of Machine Learning Research**, v. 11, p. 3011–3015, 2010.
- REN, Y. Y. et al. Predicting the aquatic toxicity mode of action using logistic regression and linear discriminant analysis. **SAR and QSAR in environmental research**, v. 27, n. 9, p. 721–746, 2016.
- RODRIGUES, L. DE B. **Ecotoxicological effects of glyphosate and formulations on different organisms**. Universidade Federal de Goiás, 2016.
- ROWAN, A. N. Ending the use of animals in toxicity testing and risk evaluation. **Cambridge quarterly of healthcare ethics : CQ : the international journal of healthcare ethics committees**, v. 24, n. 4, p. 448–458, out. 2015.
- ROY, J. et al. First report on a classification-based QSAR model for chemical toxicity to earthworm. **Journal of Hazardous Materials**, v. 386, 2020.
- SHAHRIARI, B. et al. Taking the human out of the loop: a review of Bayesian optimization. **Proceedings of the Ieee**, v. 104, n. 1, p. 148–175, 2016.
- SILVA NETO, J. A. DA. **Estudo de estabilidade de tensão em sistemas de transmissão e em microrredes utilizando o método da função energia**. 2020. Tese (Doutorado em Engenharia Elétrica) - Instituto de Sistemas Elétricos e Energia, Universidade Federal de Itajubá, Itajubá, 2020.
- UYSAL, I.; GUVENIR, H. A. Instance-based regression by partitioning feature projections. **Applied Intelligence**, v. 21, n. 1, p. 57–79, 2004.
- VAN HOUTUM, G. J. J.; VLASEA, M. L. Active learning via adaptive weighted uncertainty sampling applied to additive manufacturing. **Additive Manufacturing**, v. 48, 2021.
- VANCHINATHAN, H. P. et al. Discovering valuable items from massive data. **Kdd'15: Proceedings of the 21st Acm Sigkdd International Conference on Knowledge Discovery and Data Mining**, p. 1195–1204, 2015.
- WANG, D.; TAN, X. **Bayesian neighbourhood component analysis**. 2016.
- WANG, L. et al. Review of antidepressants in clinic and active ingredients of traditional Chinese medicine targeting 5-HT1A receptors. **Biomedicine & Pharmacotherapy**, v. 120, 2019.
- WEBER, K. C. et al. Pharmacophore-based 3D QSAR studies on a series of high affinity 5-HT 1A receptor ligands. **European journal of medicinal chemistry**, v. 45, n. 4, p. 1508–1514, 2010.
- WU, X.; ZHANG, Q.; HU, J. QSAR study of the acute toxicity to fathead minnow based on a large dataset. **SAR and QSAR in environmental research**, v. 27, n. 2, p. 147–164, 2016.
- YONDO, R.; ANDRÉS, E.; VALERO, E. A review on design of experiments and surrogate models in aircraft real-time and many-query aerodynamic analyses. **Progress in Aerospace Sciences**, v. 96, p. 23–61, 2018.

Apêndice A – Estrutura Molecular do QSAR Medicamento

Tabela A.1. Estruturas moleculares e valores de pK_i para os 81 compostos arilpiperazínicos empregados nesta investigação

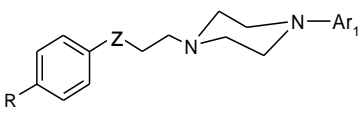
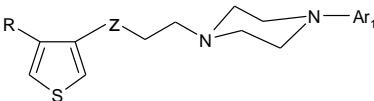
Compd.	R	Z	Ar ₁	pK_i
				
1	H	CO	2-methoxyphenyl	7.30
2	H	CHOH	2-methoxyphenyl	7.32
3	H	CHO-4-CF ₃ C ₆ H ₄	2-methoxyphenyl	6.37
4	H	CHO-4-CH ₃ OC ₆ H ₄	2-methoxyphenyl	7.04
5	H	CHO-3,4-OCH ₂ O-C ₆ H ₃	2-methoxyphenyl	7.70
6	H	CNOH	2-methoxyphenyl	7.76
7	H	CO	4-chlorophenyl	6.10
8	H	CHOH	4-chlorophenyl	6.10
9	H	CHO-4-CH ₃ OC ₆ H ₄	4-chlorophenyl	5.84
10	H	CHO-3,4-OCH ₂ O-C ₆ H ₃	4-chlorophenyl	6.26
11	H	CNOH	4-chlorophenyl	5.30
12	H	CO	4-methoxyphenyl	5.30
13	H	CHOH	4-methoxyphenyl	5.30
14	H	CHO-4-CF ₃ C ₆ H ₄	4-methoxyphenyl	5.30
15	H	CO	2-pyrimidyl	6.92
16	H	CHOH	2-pyrimidyl	6.42
17	H	CHO-4-CF ₃ C ₆ H ₄	2-pyrimidyl	5.79
18	H	CO	2-chlorophenyl	6.74
19	H	CHOH	2-chlorophenyl	6.94
20	H	CHO-4-CF ₃ C ₆ H ₄	2-chlorophenyl	5.30
21	H	CO	4-fluorophenyl	6.10
22	H	CHOH	4-fluorophenyl	6.89
23	H	CHO-4-CF ₃ C ₆ H ₄	4-fluorophenyl	5.30
24	H	CO	2-pyridyl	7.30
25	H	CHOH	2-pyridyl	6.81
26	H	CHO-4-CF ₃ C ₆ H ₄	2-pyridyl	5.79
27	H	CO	4-nitrophenyl	5.30
28	H	CHOH	4-nitrophenyl	5.30
29	H	CHO-4-CF ₃ C ₆ H ₄	4-nitrophenyl	5.30
30	phenyl	CO	2-methoxyphenyl	5.44
31	phenyl	CHOH	2-methoxyphenyl	6.07
32	phenyl	CHO-4-CF ₃ C ₆ H ₄	2-methoxyphenyl	5.30
33	methoxy	CO	2-methoxyphenyl	5.76
34	methoxy	CHOH	2-methoxyphenyl	6.49
35	methoxy	CHO-4-CF ₃ C ₆ H ₄	2-methoxyphenyl	6.00
36	nitro	CO	2-methoxyphenyl	7.30
37	nitro	CHOH	2-methoxyphenyl	8.00
				
38	H	CO	2-methoxyphenyl	7.79
39	H	CHOH	2-methoxyphenyl	7.30
40	H	CHO-4-CF ₃ C ₆ H ₄	2-methoxyphenyl	6.59
41	H	CNOH	2-methoxyphenyl	8.19
42	H	CHO-3,4-OCH ₂ O-C ₆ H ₃	2-methoxyphenyl	7.26
43	H	CHO-1-C ₁₀ H ₇	2-methoxyphenyl	6.74
44	H	CO	4-chlorophenyl	6.15

Tabela A.1. Estruturas moleculares e valores de pK_i para os 81 compostos arilpiperazínicos empregados nesta investigação

45	H	CHOH	4-chlorophenyl	5.56
46	H	CO	2-chlorophenyl	6.70
47	H	CHOH	2-chlorophenyl	6.70
48	H	CO	1-naphthyl	7.45
49	2,5-dimethyl	CO	2-methoxyphenyl	8.30
50	2,5-dimethyl	CHOH	2-methoxyphenyl	7.92
51	2,5-dimethyl	CO	2-hydroxyphenyl	8.12
52	2,5-dimethyl	CHOH	2-hydroxyphenyl	7.04
53	2,5-dimethyl	CO	1-naphthyl	7.00
54	2,5-dimethyl	CO	4-fluoro-2-methoxyphenyl	7.87
55	2,5-dimethyl	CO	4-fluoro-2-methoxyphenyl	6.30

56	H	CO	2-methoxyphenyl	8.00
57	H	CHOH	2-methoxyphenyl	7.72
58	H	CHO-1-C ₁₀ H ₇	2-methoxyphenyl	6.66
59	H	CO	4-chlorophenyl	5.30
60	H	CHOH	4-chlorophenyl	5.30
61	5-methyl	CO	2-methoxyphenyl	7.76
62	5-methyl	CHOH	2-methoxyphenyl	7.47
63	5-nitro	CO	2-methoxyphenyl	6.47

64	H	CO	2-methoxyphenyl	6.60
65	H	CHOH	2-methoxyphenyl	6.38
66	H	CO	4-chlorophenyl	5.30
67	H	CHOH	4-chlorophenyl	5.30
68	H	CO	2-hydroxyphenyl	6.00
69	H	CHOH	2-hydroxyphenyl	6.72

70	H	CO	2-methoxyphenyl	7.36
71	H	CHOH	2-methoxyphenyl	7.70
72	H	CNOH	2-methoxyphenyl	7.22
73	H	CO	4-chlorophenyl	5.30
74	H	CHOH	4-chlorophenyl	5.30
75	H	CO	2-hydroxyphenyl	6.96
76	H	CHOH	2-hydroxyphenyl	7.74
77	H	CO	4-chloro-2-methoxyphenyl	6.30
78	H	CHOH	4-chloro-2-methoxyphenyl	6.44
79	H	CO	4-fluoro-2-methoxyphenyl	6.30
80	H	CHOH	4-fluoro-2-methoxyphenyl	6.30
81	H	CO	1-naphthyl	7.00

$pK_i = -\log K_i$