

César Buchile Abud de Oliveira

**Ferramenta computacional para avaliação da
capacidade preditiva de delineamentos experimentais**

Monografia apresentada ao Instituto de Biociências, Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Botucatu, para obtenção do título de Bacharel em Física Médica.

Orientadora: Prof^ª. Dr^ª. Luzia Aparecida Trinca

Botucatu
Junho de 2014

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE - CRB 8/5651

Oliveira, César Buchile Abud.

Ferramenta computacional para avaliação da capacidade preditiva de delineamentos experimentais / César Buchile Abud Oliveira. - Botucatu, 2014

Trabalho de conclusão de curso (bacharelado - Física Médica) - Universidade Estadual Paulista, Instituto de Biociências de Botucatu

Orientador: Luzia Aparecida Trinca

Capes: 10202072

1. Delineamentos ótimos. 2. Análise de variância. 3. Processamento eletrônico de dados. 4. Física médica. 5. Experimentos clínicos - Métodos estatísticos.

Palavras-chave: DVDG; Delineamentos ótimos; Superfície de resposta; VDG; Variância de predições.

Agradecimentos

Primeiramente meus agradecimentos ao Criador, pois se não fosse pela sua graça, em todos os aspectos de minha vida, não estaria hoje concluindo este trabalho.

Aos meus pais, que acreditaram em minha capacidade e sempre me apoiaram e me orientaram dentro do que acreditavam ser o melhor para mim.

À Yasmin Ali, pelo seu apoio, pelos conselhos, pelo companheirismo e pelo carinho, o que em conjunto sempre me incentivaram.

À minha orientadora, que na verdade no decorrer deste trabalho foi muito mais que uma orientadora para mim, sempre me aconselhando não apenas nos assuntos relacionados a este trabalho, mas à minha formação, carreira e vida particular, com quem aprendi muito mais do que simplesmente realizar uma pesquisa e um relatório científico.

Aos docentes do curso de Física Médica e aos discentes da Biometria, que juntos estiveram comigo nesta caminhada, me auxiliando sempre que preciso.

Aos alunos da Física Médica Turma VIII, em especial Ariane Campolim Cristino, Isabella Paziam Fernandes Nunes, Mayara Fernandes Simões Silva, Carlos Alberto Oliveira de Biagi Júnior e Raphael de Melo Borache, pois todos tiveram uma grande contribuição na minha formação.

Ao CNPq e à Fapesp, pela concessão da bolsa de iniciação científica.

Resumo

Os delineamentos ótimos têm ganhado bastante espaço nas últimas décadas tanto do ponto de vista teórico quanto de aplicação. O uso de um delineamento ótimo num experimento, em particular, é muito importante, já que visa a obtenção do máximo de informação quando do término de um experimento. Delineamentos ótimos mais informativos têm sido construídos através do uso de critérios compostos. Uma propriedade de interesse de delineamentos é a margem de erro da predição ou estimação da resposta, quantificada pela variância da resposta predita ou estimada. Nesse projeto estudamos as propriedades preditivas de delineamentos experimentais em esquemas inteiramente aleatorizados. Essas propriedades foram estudadas através de gráficos de dispersão de variâncias (VDGs) na região experimental dos delineamentos. Para o estudo foi implementada uma rotina de otimização em linguagem C para determinar as variâncias mínima, máxima e média de predições, em hiper-esferas localizadas na região experimental. Um pacote contendo duas funções foi construído para uso no R, sendo que uma das funções utiliza a rotina em C para encontrar as variâncias máximas, mínimas e médias e a outra utiliza os resultados fornecidos por esta para a elaboração do VDG.

Sumário

1	Introdução	5
2	Fundamentação teórica	6
2.1	A aproximação da superfície de resposta por um polinômio	6
2.1.1	Modelos de regressão linear	7
2.2	A variância das predições	10
2.3	A relação de uma superfície de resposta com o VDG	11
2.4	Funções critério para construção de delineamentos ótimos	13
3	Metodologia	14
3.1	Programa em C	14
3.2	Construção do pacote instalável no R em plataforma <i>Windows</i>	16
4	Resultados e Discussões	17
4.1	Exemplo 1 - Comparação dos delineamentos Box-Behnken, Central Composto e o obtido por meio do critério D	17
4.2	Exemplo 2 - Comparação dos delineamentos obtidos através dos critérios D_S , DP_S e Composto	19
5	Conclusão	20
A	Delineamentos comparados para o Exemplo 1	22
B	Funções do pacote <i>Dispersion</i>	23

1 Introdução

Em diversas áreas da ciência a estatística é utilizada como ferramenta de planejamento de coleta, bem como de análise e interpretação, tanto para estudos experimentais como observacionais. Uma pesquisa experimental bem planejada e controlada é considerada a forma mais poderosa de pesquisa empírica, evitando-se assim custos desnecessários e experimentos fornecendo resultados não informativos.

O planejamento de um experimento envolve o levantamento das condições que podem interferir na resposta a ser investigada, usualmente referidas como fatores experimentais. Nesta fase, o pesquisador deve escolher os fatores experimentais que deseja investigar, decidir como estes serão variados, bem como dimensionar o tamanho da amostra que irá utilizar, sendo este definido como tamanho do experimento ou número de unidades experimentais.

Os fatores podem ser quantitativos ou qualitativos, mas neste trabalho será considerado experimentos nos quais todos os fatores são quantitativos contínuos. Os níveis máximo e mínimo de cada fator definem a região experimental, na qual cada ponto é um tratamento, ou seja, uma combinação dos níveis dos fatores. Note que para fatores contínuos a região é compacta e contém infinitos pontos ou tratamentos.

Experimentos clássicos utilizam os tratamentos do fatorial completo resultante da escolha de níveis igualmente espaçados para cada fator, como por exemplo o fatorial completo para 3 fatores, cada um com 3 níveis, o qual resulta em 3^3 tratamentos distintos.

Realizar um experimento utilizando o fatorial completo por vezes é inviável tanto do ponto de vista prático quanto econômico, sendo assim necessário, de alguma maneira, escolher um subconjunto de pontos dentro do fatorial, ou da região experimental, visando a seleção do subconjunto mais informativo com relação ao problema estudado. Tal subconjunto de pontos selecionados, aliado às regras de distribuição dos tratamentos às unidades experimentais, caracteriza o delineamento experimental. Logicamente, para uma região experimental existe um grande número de delineamentos alternativos que fornecerão boas propriedades a respeito do problema estudado, mas alguns produzirão resultados melhores que outros, sendo assim necessário compará-los. Com o avanço computacional percebeu-se que um delineamento informativo poderia ser escolhido via um algoritmo de otimização, uma vez que uma função critério seja definida. Uma função critério considera uma certa propriedade de interesse do delineamento que deve ser otimizada. Em Box e Draper (1987) e mais recentemente em Atkinson, Donev e Tobias (2007) foram apresentadas várias funções critério que podem ser utilizadas, no entanto estas são unidimensionais, e por vezes o interesse está em propriedades multidimensionais.

Quando o objetivo do experimento é prever ou estimar a resposta na região experimental, o estudo da variância da predição, com domínio em tal região, desempenha papel importante na escolha do melhor delineamento.

Dada a expressão do modelo estatístico que se deseja ajustar para compreender as relações entre a resposta e os fatores, é possível investigar o comportamento da função variância na região experimental antes de realizar o experimento. No entanto, esta função é multidimensional, sendo assim necessária alguma simplificação ou obtenção de medidas resumidas. O VDG (*Variance Dispersion Graph*), proposto em Jensen e Myers (1989), desempenha tal papel quando o modelo visado é um polinômio de primeiro ou de segundo grau e a região experimental é do tipo cuboidal ou esférica. Para tanto, um gráfico é elaborado utilizando os valores calculados para as variâncias máximas, mínimas e médias da resposta predita em função da distância dos pontos em relação ao centro da região experimental. Em suma, para a elaboração de um VDG é então necessário encontrar as variâncias máximas, mínimas e médias em várias hiperesferas circunscritas dentro da região experimental escolhida. Um programa em linguagem Fortran para os cálculos necessários foi implementado em Vining (1993).

Como a linguagem C é considerada mais atual e utilizada, em comparação com a Fortran, além de ser muito indicada para rotinas iterativas, o objetivo deste trabalho foi implementar em linguagem C o procedimento apresentado em Vining (1993), vislumbrando uma conexão entre o C e o programa estatístico R através da instalação de um pacote. Para tanto, trataremos todo o conteúdo teórico necessário para o bom entendimento do programa implementado em Vining (1993) e apresentaremos os resultados obtidos com seu equivalente implementado em C. Abordaremos também a estrutura de um pacote do R e o procedimento necessário para a construção de um pacote contendo funções em R que se utilizam de uma rotina em C. Buscando demonstrar os resultados obtidos, apresentaremos e discutiremos VDGs elaborados para alguns delineamentos utilizando o pacote construído.

2 Fundamentação teórica

Nesta seção apresentamos uma introdução às metodologias de superfície de resposta, a definição de variância de predições, uma introdução ao conceito de delineamentos ótimos juntamente com as funções critério mais utilizadas para construí-los.

2.1 A aproximação da superfície de resposta por um polinômio

De forma geral, um experimento tem por objetivo, quando do seu término, obter dados que permitam ao pesquisador estudar e tirar conclusões fidedignas a respeito de uma certa entidade estudada, a qual usualmente é denominada por variável resposta. Normalmente tal estudo é realizado ajustando-se um polinômio às observações da variável resposta, sendo este regido pelos fatores experimentais utilizados. Os modelos polinômiais pertencem à ampla classe dos modelos lineares ou, mais especificamente, dos modelos de

regressão linear múltipla, assunto tratado com maiores detalhes na seção 2.1.1.

Quando elaboramos uma representação gráfica do ajuste polinomial realizado em função dos fatores experimentais que o regem, obteremos uma curva, se apenas um fator estiver sendo utilizado, ou uma superfície em dimensão uma unidade maior que o número de fatores utilizados, quando dois ou mais fatores estiverem sendo usados. A esta superfície gerada denominamos por superfície de resposta, a qual permite estudar como os fatores experimentais interferem no comportamento da variável resposta.

Contudo, sabe-se que um ajuste polinomial raramente englobará todos os pontos interpolados, havendo assim uma diferença entre os valores de fato observados no experimento em relação aos preditos pelo polinômio. Assim, todas as predições feitas utilizando a superfície de resposta obtida terão uma margem de erro, e quanto menor for a mesma mais confiáveis serão as predições. Por conseguinte, quando temos delineamentos experimentais alternativos, é recomendado escolher o que fornecerá uma margem de erro menor, resultando assim em predições mais confiáveis. Devido à margem de erro ser definida como a raiz quadrada da variância, todo o estudo de escolha pode, sem perda de generalidade, ser realizado na escala das variâncias.

Por fim, uma ferramenta útil que permite o estudo da variância das predições para um dado delineamento é o VDG, assunto que será melhor abordado nas seções 2.2 e 2.3.

2.1.1 Modelos de regressão linear

Considere uma variável X que tenha alguma relação com uma variável Y . Tal situação sugere uma maneira alternativa de se estudar Y , bastando para isso incorporar, na equação de Y , as informações sobre X . A ideia mais simples é a especificação do **Modelo de Regressão Linear Simples (MRLS)**, no qual Y é uma variável aleatória de interesse (variável resposta) e X uma variável (não necessariamente aleatória) denominada **auxiliar, explicativa ou regressora**. Em experimentos, as variáveis regressoras são os fatores experimentais utilizados e as suas interações. Têm-se uma descrição da variável Y como uma soma de quantidades determinísticas (sobre uma reta em função de X , por exemplo) e uma quantidade aleatória (inúmeros fatores que podem, conjuntamente, interferir em Y), ou seja,

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad (1)$$

na qual β_0 e β_1 são os coeficientes da reta e X assume valores considerados fixos. Em tal modelo, é suposto que a esperança de ϵ seja igual a zero, bem como que sua variância seja independente do valor específico de X . Matematicamente, escrevemos: $E[\epsilon] = 0$; $\text{Var}[\epsilon] = \sigma_\epsilon^2$.

Tais teorias, na prática, são aplicadas tomando-se uma amostra de n unidades de uma população de interesse, das quais toma-se as medidas (x, y) , de tal maneira que o modelo

de regressão linear é agora representado por:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n \quad (2)$$

sendo cada quantidade aleatória independente de uma outra qualquer. Isso, uma vez suposto a distribuição normal, estatisticamente é representado em forma de covariância identicamente a zero, ou seja, $\text{Cov}[\epsilon_i, \epsilon_j] = 0$, $i, j = 1, \dots, n$ para $i \neq j$.

O modelo em (2) é escrito em forma matricial por:

$$\begin{aligned} \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} &= \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\ &= \begin{bmatrix} \beta_0 + \beta_1 x_1 + \epsilon_1 \\ \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \epsilon_n \end{bmatrix} \\ \mathbf{Y} &= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \end{aligned}$$

na qual a matriz \mathbf{X} é denominada matriz do modelo, com número de colunas igual ao número de elementos de $\boldsymbol{\beta}$ e número de linhas igual ao tamanho da amostra. A coluna unitária é referente ao intercepto, o elemento que jamais varia no modelo. O vetor aleatório $\boldsymbol{\epsilon}$, na situação mais simplificada, é composto por variáveis independentes, com distribuição $N(\mathbf{0}; \sigma_\epsilon^2 \mathbf{I})$, tendo assim vetor de esperança nulo e matriz de variâncias e covariâncias, de dimensão $n \times n$, dada por:

$$\text{Var}[\boldsymbol{\epsilon}] = \begin{bmatrix} \sigma_\epsilon^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_\epsilon^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_\epsilon^2 \end{bmatrix} = \sigma_\epsilon^2 \mathbf{I},$$

na qual a diagonal é formada pelas variâncias e os demais elementos são as covariâncias entre pares de erros, consideradas nulas. A constância dos elementos da diagonal significa que a variância não depende dos valores de X e essa propriedade é chamada de homocedasticidade ou homogeneidade de variâncias. Assim, tem-se que:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{com } \boldsymbol{\epsilon} \sim NM_n(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}). \quad (3)$$

O modelo em (3) é um polinômio de primeira ordem, com um fator, e dependendo da relação entre variáveis resposta e explicativa, pode ser necessário aumentar a ordem do polinômio, acrescentando termos quadráticos e outras variáveis explicativas, bem como interações entre estas. Nestes casos o vetor β tem dimensão maior do que dois, e a matriz \mathbf{X} é definida de acordo com as suposições do modelo que está sendo estudado.

Quando a matriz \mathbf{X} tem mais de duas colunas o modelo é chamado de **Modelo de Regressão Linear Múltiplo (MRLM)**, cujo caso mais simples também assume todas as premissas em relação ao erro já feitas para o **MRLS**.

Assim como no **MRLS**, o **MRLM** também pode ser escrito em forma matricial,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

com a mesma estrutura de erros do modelo em (3), em que

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix},$$

com o número de colunas de \mathbf{X} igual ao número de elementos de β e o número de linhas de \mathbf{X} igual ao tamanho da amostra para cada **MRLM** apresentado. O índice p indica número de parâmetros na equação associados às variáveis regressoras. Pode-se também ter uma extensão das colunas de \mathbf{X} levando-se em consideração interações entre variáveis regressoras ou até mesmo, conforme já mencionado, a inclusão de colunas para incorporarem termos quadráticos e demais ordens na equação, tudo dependendo apenas das suposições feitas para as relações entre Y e as variáveis regressoras X 's.

Para a resolução do sistema referente ao **MRLM** buscando encontrar estimativas para β , $\hat{\beta}$, é utilizado o método dos mínimos quadrados para os erros, buscando minimizar o somatório das diferenças quadráticas entre o valor de y correspondente e o valor dado pela função linear em β , ou seja, a parte preditiva do modelo. Sendo assim, busca-se minimizar a expressão $(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$ estimando β pelo método de mínimos quadrados. Resolvendo o problema de minimização chega-se na expressão para o estimador de β ,

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \quad (4)$$

em que \mathbf{X}' é a notação utilizada para indicar a matriz transposta de \mathbf{X} . Note que o estimador de β só é único se $(\mathbf{X}'\mathbf{X})$ for invertível, ou seja, se \mathbf{X} for de posto completo. Este estimador é uma combinação linear de \mathbf{Y} e suas propriedades estatísticas como

esperança e variância são dadas por, respectivamente:

$$E[\widehat{\boldsymbol{\beta}}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = \boldsymbol{\beta}, \quad (5)$$

e

$$\text{Var}[\widehat{\boldsymbol{\beta}}] = \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = \sigma_{\epsilon}^2(\mathbf{X}'\mathbf{X})^{-1}. \quad (6)$$

Note que σ_{ϵ}^2 é a variância dos erros e não depende dos dados observados e nem de \mathbf{X} . Sendo assim, a variância em (6) depende de uma parte sob controle do pesquisador, \mathbf{X} , e de uma parte que independe do experimento, σ_{ϵ}^2 , sendo este último obtido pelo estimador de mínimos quadrados,

$$\widehat{\sigma}_{\epsilon}^2 = \frac{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{n - p - 1}, \quad (7)$$

ao término do experimento.

2.2 A variância das predições

Conforme descrito na seção 2.1, geralmente o pesquisador, quando conclui seu experimento, obtém uma superfície de resposta que o auxilia em suas inferências e conclusões relacionadas à variável resposta. Tal superfície é obtida através de modelos de regressão linear, os quais podem ser tanto um MRLS quanto um MRLM. Dado o modelo ajustado, uma predição da resposta no ponto $\mathbf{x} \in \mathbf{R}$, sendo \mathbf{R} a região experimental, é dada por

$$\widehat{y}(\mathbf{x}) = \mathbf{f}'(\mathbf{x})\widehat{\boldsymbol{\beta}}, \quad (8)$$

cuja variância é

$$\text{Var}[\widehat{y}(\mathbf{x})] = \sigma_{\epsilon}^2(\mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f}(\mathbf{x}) + 1), \quad (9)$$

e $\mathbf{f}(\mathbf{x})$ é uma função que expande o vetor \mathbf{x} de maneira a gerar um vetor com todos os termos do modelo polinomial de interesse, sendo $\mathbf{f}'(\mathbf{x})$ o vetor transposto do vetor expandido por $\mathbf{f}(\mathbf{x})$. Já a estimativa da **resposta esperada** no ponto $\mathbf{x} \in \mathbf{R}$, ou seja $\widehat{E}[y(\mathbf{x})]$, é a mesma dada em (8), no entanto com variância menor, dada por

$$\text{Var}[\widehat{E}[y(\mathbf{x})]] = \sigma_{\epsilon}^2\mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f}(\mathbf{x}). \quad (10)$$

Note que a diferença entre as expressões (9) e (10) é uma constante e assim, comparamos delineamentos distintos com base na margem de erro de predição pela fórmula da variância apresentada em (10).

Como σ_ϵ^2 em (10) é uma constante, podemos estudar

$$\frac{\text{Var}[E[\hat{y}(\mathbf{x})]]}{\sigma_\epsilon^2} = \mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f}(\mathbf{x}), \quad (11)$$

ou assumir $\sigma_\epsilon^2 = 1$, sem perda de generalidade. Na equação (11) fica clara a ideia de que a $\text{Var}[E[\hat{y}(\mathbf{x})]]$ depende exclusivamente da matriz de delineamento e do modelo a ser ajustado, e por assim ser, é possível trabalharmos com tal medida antes mesmo de executarmos o experimento, uma vez que σ_ϵ^2 é uma constante e pode ser estimada ao término deste por (7).

Em geral um experimento não consegue prever as respostas com a mesma variância em toda região experimental, existindo delineamentos com capacidade de fazer boas previsões no centro, nas bordas, ou até mesmo em outras partes da região experimental. Todavia, é muito difícil um delineamento ser eficiente em toda a região. Existem critérios para a construção de delineamentos que minimizam a variância máxima ou a variância média, no entanto os delineamentos resultantes não necessariamente apresentarão boa capacidade preditiva na porção mais importante para o pesquisador. Na prática, dado delineamentos distintos, podemos estudar a função da variância das previsões e fazer uma escolha qualitativa de qual deles recomendar para uso.

2.3 A relação de uma superfície de resposta com o VDG

Quando pensamos em escolher o melhor delineamento em um conjunto de vários delineamentos, sendo cada um obtido por métodos diferentes, buscamos selecionar o que apresenta capacidade preditiva mais adequada. Desta maneira, é compreensivo utilizarmos as variâncias de previsões de todos os delineamentos como ferramenta de comparação. Não obstante, o estudo das variâncias de previsões teria de ser feito ponto a ponto em toda região experimental e de alguma maneira escolhermos o que em geral apresentasse variâncias de previsões menores.

Nos casos em que um ou dois fatores são utilizados, poderíamos gerar as curvas, ou superfícies, resultantes da função variância da previsão em função do delineamento e do modelo de regressão linear utilizado, conforme equação (11). Em seguida, uma comparação visual entre as curvas obtidas, ou superfícies, poderia ser feita. Contudo, em delineamentos com mais de dois fatores experimentais, a comparação visual se torna impossível devido à alta dimensão.

Jensen e Myers (1989) apresentaram uma alternativa para lidar com a alta dimensionalidade introduzindo a metodologia do gráfico de dispersão da variância (VDG), a qual resume a informação relacionada à superfície gerada pelas variâncias de previsões em dados que possibilitam seu estudo sempre em curvas elaboradas em duas dimensões, independentemente do número de fatores utilizados pelo delineamento.

Posteriormente, os cálculos que envolvem tal metodologia foram implementados em linguagem Fortran em Vining (1993). Nos parágrafos que se seguem nesta seção serão apresentados os conceitos relacionados à metodologia do VDG conforme apresentado em Jensen e Myers (1989).

A variância da predição para um delineamento em uma região experimental não depende exclusivamente da distância do tratamento estudado até o centro do delineamento, salvo se o delineamento for do tipo rotacional. A definição de **delineamento rotacional** é exatamente esta, a variância da predição é a mesma para todos os pontos equidistantes do centro da região experimental.

Para uma esfera (ou hiperesfera) de raio r , a **variância esférica média** de uma predição, é definida por

$$V^r = \frac{\Psi}{\sigma_\epsilon^2} \int_{U_r} \text{Var}[\hat{y}(\mathbf{x})] d\mathbf{x}, \quad (12)$$

em que $U_r = \left\{ \mathbf{x} : \sum_{i=1}^k x_i^2 = r^2 \right\}$ e $\Psi^{-1} = \int_{U_r} d\mathbf{x}$ é a área, volume ou hipervolume da superfície de U_r .

Resolvendo a integral em (12) e usando o fato de V^r ser um escalar, obtemos

$$V^r = \text{tr}(\mathbf{S}(\mathbf{X}'\mathbf{X})^{-1}), \quad (13)$$

na qual \mathbf{S} é a matriz de momentos da região. Para regiões esféricas, \mathbf{S} pode ser escrita como:

$$\mathbf{S} = \Psi \int_{U_r} \mathbf{f}(\mathbf{x})\mathbf{f}'(\mathbf{x}) d\mathbf{x}. \quad (14)$$

Desta maneira, dado o modelo, os momentos são funções do raio r e de Ψ , a dimensão da hiperesfera. Para delineamentos rotacionais, o estudo da capacidade preditiva do delineamento pode ser feito elaborando gráficos da variância esférica média em função do raio.

Quando o delineamento não é rotacional é interessante avaliar a instabilidade, definida na equação (15), bem como a ordem de magnitude dos valores da função variância de predições.

Primeiramente, devemos construir um gráfico de $\text{Var}[\hat{y}(\mathbf{x})]$ (ou de $n\text{Var}[\hat{y}(\mathbf{x})]$, se delineamentos de tamanhos diferentes, n , estiverem sob comparação) em função do raio. A instabilidade pode ser representada como

$$\max_{\mathbf{x} \in U_r} \left[\frac{n\text{Var}[\hat{y}(\mathbf{x})]}{\sigma_\epsilon^2} \right] - \min_{\mathbf{x} \in U_r} \left[\frac{n\text{Var}[\hat{y}(\mathbf{x})]}{\sigma_\epsilon^2} \right]. \quad (15)$$

Para a elaboração do VDG, um gráfico é elaborado utilizando as variâncias máximas, mínimas e médias calculadas sobre várias hipersferas concêntricas localizadas na região do delineamento, em função do raio da hipersfera a que pertencem. Tal gráfico possibilita uma comparação visual da instabilidade de vários delineamentos, além das variâncias das predições como um todo à medida que os tratamentos se afastam do centro do delineamento, o qual, *a priori*, é considerado o ponto que produzirá a melhor resposta.

2.4 Funções critério para construção de delineamentos ótimos

As funções critério, como já comentado anteriormente, foram propostas com o intuito de auxiliar na obtenção de um delineamento dito ótimo, isto é, obtido através da escolha de um conjunto de pontos, ou tratamentos, pertencentes a uma região experimental, através da otimização de uma função matemática que leva em conta alguma característica específica do delineamento que está sendo formado, a qual definimos por função critério. Em suma, a construção do delineamento ótimo busca $\mathbf{X} \in \mathbf{R}$ tal que alguma propriedade de \mathbf{X} seja otimizada, isto é, de acordo com a função critério escolhida. Algumas das funções critério mais utilizadas são

1. $|(\mathbf{X}'\mathbf{X})^{-1}|$, chamado de critério D ;
2. $F(p; d; 1 - \alpha)|(\mathbf{X}'\mathbf{X})^{-1}|$, chamado de critério DP ;
3. $tr((\mathbf{X}'\mathbf{X})^{-1})$, chamado de critério A ;
4. $tr(\mathbf{W}(\mathbf{X}'\mathbf{X})^{-1})$, chamado de critério A ponderado, no qual \mathbf{W} é um vetor de pesos;
5. $F(1; d; 1 - \alpha)tr\{\mathbf{W}(\mathbf{X}'\mathbf{X})^{-1}\}$, chamado de critério AP ;
6. $\int_{\mathbf{x} \in \mathbf{R}} \mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f}(\mathbf{x})d\mathbf{x}$, chamado de critério I ;
7. $\max_{\mathbf{x} \in \mathbf{R}} [\mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f}(\mathbf{x})]$, chamado de critério G .

Em todos os casos quanto menor o valor da função critério melhor o delineamento. O critério DP foca no teste F global para todos termos do modelo e a quantidade $F(p; d; \alpha)$ é definida como o quantil de ordem $(1 - \alpha)$ da distribuição F de *Snedecor* com p (número de variáveis regressoras no modelo) graus de liberdade no numerador e $d = n - T$ graus de liberdade no denominador devido a T tratamentos distintos no delineamento, ou seja, graus de liberdade do erro puro, e α é o nível de significância do teste. Já o critério AP foca em testes ou intervalos de confiança individuais para cada parâmetro no modelo e $F(1; d; 1 - \alpha)$ é o respectivo quantil da distribuição F . Para as funções critério apresentadas nos itens de 1 a 5, existem também as versões que dão ênfase a um sub-conjunto do conjunto de parâmetros, ou seja, alguns parâmetros são considerados do tipo *nuisance*.

Nesses casos, as funções critério são chamadas pela mesma letra acrescidas do índice S (exemplos: critério D_S , AP_S).

Porém, cada um dos critérios apresentados considera uma única propriedade do delineamento. Um delineamento ótimo com relação a uma propriedade pode ser sub-ótimo em relação a outras. Como na prática um experimento tem diversos objetivos, a combinação de várias propriedades numa única função critério é bastante atrativa. A função que combina os vários critérios é chamada de **critério composto**. A função que rege o critério composto, proposto em Trinca e Gilmour (2012), é definida como

$$\frac{|\mathbf{X}'_R \mathbf{Q}_0 \mathbf{X}_R|^{\frac{\kappa_1}{(p-1)}} (n-d)^{\kappa_4}}{\{F(p-1; d; 1-\alpha_1)\}^{\kappa_1} \{F(1; d; 1-\alpha_2)\}^{\kappa_2} \{tr[\mathbf{W}(\mathbf{X}'_R \mathbf{Q}_0 \mathbf{X}_R)^{-1}]\}^{\kappa_2+\kappa_3}}, \quad (16)$$

no qual \mathbf{Q}_0 é a matriz responsável pela obtenção da matriz de informação de $\boldsymbol{\beta}$ excluindo-se o intercepto, κ_i é o peso relacionados ao critério i , sendo κ_1 relacionado com o critério DP , κ_2 com o critério AP , κ_3 com o critério A e κ_4 com os graus de liberdade. O termo \mathbf{X}_R é a matriz \mathbf{X} sem a coluna referente ao intercepto. Os pesos referentes a κ_i devem ser escolhidos de modo a refletir a importância relativa dos diferentes aspectos da análise, podendo inclusive, em alguns casos, ser nulo.

Sabemos também que \mathbf{Q}_0 é obtido por $\mathbf{I} - \bar{\mathbf{J}}$, em que $\bar{\mathbf{J}} = \frac{1}{n} \mathbf{1}\mathbf{1}'$, bem como que a matriz de informação para este caso em que considera-se a matriz \mathbf{X} sem o intercepto é dada por $(\mathbf{X}'_R(\mathbf{I} - \bar{\mathbf{J}})\mathbf{X}_R)$, assim temos que a variância para o vetor $\hat{\boldsymbol{\beta}}_S$ definido como sendo o sub-conjunto de estimadores de $\boldsymbol{\beta}$ excluindo-se o termo $\hat{\beta}_0$ é dada por:

$$\text{Var}[\hat{\boldsymbol{\beta}}_S] = (\mathbf{X}'_R(\mathbf{I} - \bar{\mathbf{J}})\mathbf{X}_R)^{-1}, \quad (17)$$

para $\sigma_\epsilon^2 = 1$, sem perda de generalidade.

3 Metodologia

Nesta seção apresentamos as principais subrotinas que constituem a rotina em C, bem como a função de cada uma delas no cálculo da variância de predições máxima, mínima e média. O conceito de um pacote no R é apresentado sucintamente, assim como o procedimento utilizado para a construção do pacote *Dispersion*.

3.1 Programa em C

Para a construção do gráfico de dispersão da variância (VDG) para um experimento inteiramente aleatorizado, usamos a rotina em C que calcula as variâncias de predições apresentada na equação (11), buscando os valores mínimo e máximo sobre a superfície da hipersfera de raio r . Variando-se o valor de r , com $0 \leq r \leq 1$, percorre-se toda a

região experimental. O número de hiperesferas em que a busca será realizada é escolhido pelo usuário. Dos valores calculados, o valor máximo e o mínimo são guardados para uso posterior.

No processo de otimização, a subrotina *MINLOC()* é a responsável por encontrar o valor máximo e mínimo da variância da predição, e conta com o auxílio da subrotina *AMOEBAS()*, uma implementação do método *Simplex* de *Nelder-Mead*, e com uma busca em grade, sendo a grade inicial gerada pela subrotina *GRID()*.

Já a variância média é encontrada pelo cálculo da equação (13), sendo a matriz de momentos formada através do cálculo da integral apresentada na equação (14). No entanto, para os modelos polinomiais os momentos têm um padrão sistemático, sendo função do raio da hiperesfera e do número de fatores, como apresentado em Jensen e Myers (1989) para delineamentos em região experimental esférica. Desta maneira, a subrotina *VSPH()* calcula as variâncias médias para todos os raios utilizando a matriz de momentos. Para delineamentos em região experimental cuboidal a matriz de momentos utilizada é a mesma que para delineamentos em região esférica, logo para os raios que ultrapassam a região do delineamento, as variâncias médias obtidas não estão corretas, não sendo aconselhável utilizá-las.

A subrotina *PROCV()* tem por objetivo gerenciar tanto a subrotina *MINLOC()* quanto a subrotina *VSPH()*. A equação (11) é calculada através da subrotina *EVAL()*.

A matriz de delineamento é expandida ao modelo de primeira ou segunda ordem completo com o auxílio da subrotina *EXPAND()*, responsável por expandir a matriz de delineamento linha por linha segundo o modelo de regressão linear escolhido pelo usuário. Após a multiplicação da matriz de delineamento expandida pela sua transposta, esta é invertida com o auxílio da subrotina *INVERT()* pela resolução de sistemas por decomposição LU, sendo as subrotinas responsáveis por tal a *LUBKSB()* e a *LUDCMP()*, ambas obtidas em Press *et al.* (1992).

Quando a região experimental é cuboidal, devido às buscas otimizadas serem realizadas sobre hiperesferas, após um dado raio estas necessitam ser restringidas para que os cálculos sejam realizados apenas nas regiões localizadas dentro do hiper-cubo.

Uma vez concluída a busca até o raio unitário, isto é, para todas as hiperesferas escolhidas, as variâncias máximas, mínimas e médias, em função do raio da hiperesfera que pertencem, são retornadas como uma matriz no *workspace* do R. Tais dados são então utilizados para a elaboração do VDG.

Por exigência da estrutura da interação de programas em C com o R, a rotina principal implementada em C teve de receber um nome diferente do usual *main()*, sendo usado o nome *principal()* e declarada como tipo *void*. Todas as subrotinas foram declaradas dentro da rotina *principal()*. Todos os dados de entrada são transferidos para a rotina *principal()* como argumentos sob forma de ponteiros. A impressão de dados no prompt do R é realizada com a função *Rprintf()*. As bibliotecas *R.h* e *Rmath.h* foram adicionadas

com o comando *include*.

A implementação foi realizada em plataforma *Windows* com o auxílio do programa *Dev-C++*.

3.2 Construção do pacote instalável no R em plataforma *Windows*

O pacote desenvolvido para o R, denominado *Dispersion*, conta com duas funções, a *Variance.dispersion()*, a qual se utiliza da rotina em C para encontrar as variâncias máximas, mínimas e médias para vários raios e as disponibilizar em uma matriz no R, e a *Plot.dispersion()*, utilizada para plotar os VDGs para até dois delineamentos.

Um pacote no R é definido como uma coletânea de funções e estruturas de dados que são adequados à abordagem de um problema particular, sendo facilmente distribuído e compartilhado com outras pessoas.

Uma vez escrito e devidamente testado, o pacote pode ser disponibilizado em três principais repositórios para pacotes do R: o *CRAN*, o *Bioconductor* e o *Omegahat*, sendo o primeiro o mais conhecido e utilizado. Uma vez disponibilizado o pacote em um dos repositórios, ele pode facilmente ser consultado por qualquer usuário, copiado e instalado, estando assim pronto para uso.

Um pacote no R é estruturado basicamente em diretórios que contêm as funções em linguagem R, os arquivos de ajuda, dados, códigos escritos em outras linguagens, e demais arquivos necessários para o pleno funcionamento do mesmo.

Para construir o pacote, utilizamos o programa R-3.0.0 e o Rtools-3.0, sendo o segundo um conjunto de ferramentas necessárias para a construção de um pacote. Os passos realizados para a criação do pacote foram:

1. O *path*, em variáveis de ambiente do sistema, foi configurado de maneira a incluir o caminho para os arquivos executáveis tanto do R quanto das ferramentas do Rtools.
2. As funções *Variance.dispersion()* e a *Plot.dispersion()* foram escritas em um script. sendo ambas carregadas no prompt de comando do R, juntamente com alguns delineamentos para a posterior elaboração de exemplos nos arquivos de ajuda.
3. A função *package.skeleton()* foi invocada no prompt de comando do R, visando criar um diretório inicial para o pacote *Dispersion*. Tal diretório contém arquivos com os requisitos básicos para um pacote válido no R.
4. Edições nos arquivos gerados no passo anterior foram realizadas com o auxílio de um editor de textos.
5. O diretório do pacote *Dispersion* foi acessado e criado dentro deste um subdiretório de nome *scr*, sendo neste inserido o código fonte da rotina *principal()*, escrita em C.

6. Um arquivo, nomeado como *NAMESPACE*, foi criado no diretório do pacote *Dispersion* usando um editor de textos. Neste os comandos *useDynLib(Dispersion)* e *export(Variance. dispersion, Plot.dispersion)*, foram inseridos, responsáveis por disponibilizar as funções contidas no pacote ao usuário do R quando este inicia o mesmo.
7. No prompt de comando do Windows foi digitado o comando *R CMD check Dispersion* para conferir possíveis erros e avisos estruturais existentes no pacote. Após sanar todos eles, executamos o comando *R CMD INSTALL - -build Dispersion* para criar um arquivo compactado instalável no R.
8. No menu *Packages* do R, a opção *Install package(s) from local zip files...* foi acessada para instalar o pacote. Em seguida, no mesmo menu foi acessada a opção *Load package...* e selecionado o pacote *Dispersion* na lista que apareceu. Para utilizá-lo, basta digitar o comando *library(Dispersion)* no prompt de comando do R.

Para invocar a rotina em C dentro da função *Variance.dispersion()* utilizamos o comando *.C("principal", variáveis passadas para a rotina)*. O comando *.C()* retorna uma lista com as variáveis originais e/ou modificadas pela rotina em C, sendo necessário atribuir esta a uma variável para posterior uso. As variáveis passadas para a rotina *principal()* tiveram que ser classificadas de acordo com sua declaração na mesma (*as.integer(variável)*, *as.double(variável)* e etc).

Maiores informações a respeito da construção de pacotes no R, bem como suas estruturas, podem ser encontrados em Chambers, Hand e Härdle (2008) e Gentleman (2009).

4 Resultados e Discussões

Nesta seção exemplificamos a elaboração dos VDGs por meio do pacote *Dispersion* para dois exemplos. No primeiro, utilizamos o delineamento clássico Box-Behnken usado em um experimento descrito em Li *et al.* (2013) e o comparamos com outros dois delineamentos, a saber o Central Composto e um obtido pelo critério *D* (Anexo 1). Já no segundo, delineamentos apresentados na Tabela 2 de Trinca e Gilmour (2012) foram comparados com base nos VDGs que os originaram.

4.1 Exemplo 1 - Comparação dos delineamentos Box-Behnken, Central Composto e o obtido por meio do critério D

Neste exemplo utilizamos um delineamento Box-Behnken (ver Box e Draper (1987)) com 5 fatores, cada um com três níveis, e 46 tratamentos, utilizado em Li *et al.* (2013) para encontrar o tratamento ótimo responsável por acelerar a separação do complexo

extrato bruto de folhas de *Rosa sericea* através da cromatografia por fluido supercrítico de alta eficiência, a qual torna possível, por exemplo, a separação de substâncias não voláteis em colunas abertas, visto que um fluido supercrítico reúne a vantagem da alta difusibilidade do gás e do alto poder de solvatação do líquido. Todos os fatores estão relacionados com as condições de eluição (X_1 : tempo gradiente T_I (min); X_2 : relação da fase móvel de metanol B (I) em % para o gradiente do primeiro solvente; X_3 : tempo gradiente T_{II} (min); X_4 : relação da fase móvel de metanol B (II) em % para o gradiente do segundo solvente; X_5 : concentração de TFA).

As folhas de *Rosa sericea* são largamente utilizadas na medicina tradicional chinesa por ter propriedades anti-*Helicobacter pylori*, anti-HIV, anti-bactericida, anti-oxidante, eliminar radicais livres, dentre outras, ficando assim justificada a importância de separar os vários componentes presentes nas folhas de *Rosa sericea* para melhor estudá-los e entender seus princípios ativos.

Elaboramos três VDGs neste exemplo, sendo um deles para o delineamento Box-Behnken utilizado em Li *et al.* (2013), e os outros dois para um delineamento obtido otimizando o critério D e um para o Central Composto (ver Box e Draper (1987)), ambos com o mesmo número de fatores e níveis que o Box-Behnken. A região experimental utilizada para os cálculos e elaboração dos VDGs foi a cuboidal, e o modelo de regressão linear utilizado foi o de segunda ordem completo, isto é, incluindo todas as interações duplas entre os fatores. Os delineamentos Box-Behnken e Central Composto são clássicos no estudo de superfície de respostas. Os VDGs obtidos estão apresentados na Figura 1.

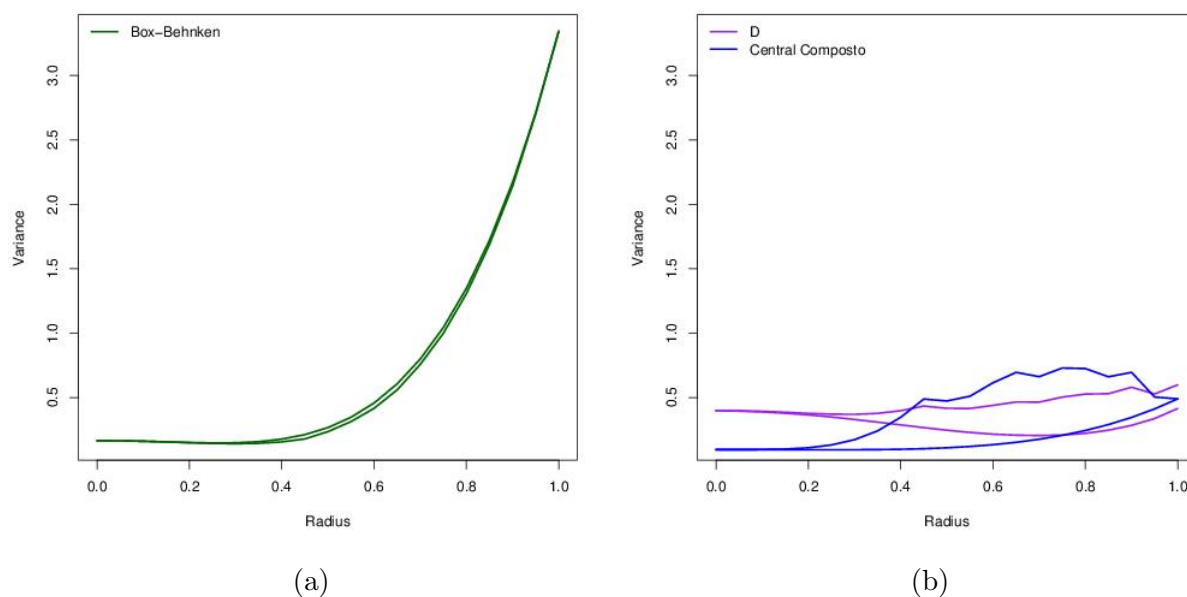


Figura 1: VDGs do delineamento utilizado em Li *et al.* (2013) e outros dois delineamentos alternativos, sendo um o Central Composto e o outro obtido pelo critério D .

Estudando os VDGs apresentados na Figura 1 verificamos que o delineamento Box-

Behnken apresentou comportamento próximo de um rotacional, com variâncias, até por volta do raio 0,4, bem estáveis, no entanto atinge valores crescentes a partir deste valor, e estes são maiores que para os dois outros delineamentos estudados. Logo, é considerado como tendo boa capacidade preditiva para combinações dos fatores mais próximos do centro da região experimental. O delineamento Central Composto é o melhor dentre todos os apresentados por predizer respostas no centro e nas extremidades da região experimental. O delineamento D é o que mais controla as variâncias em média, possuindo melhor capacidade preditiva nas extremidades da região experimental. Não obstante, deixa a desejar com relação ao Central Composto, devido às variâncias, inclusive as mínimas, relativamente altas em todo o espectro de raios até 0,4.

Assim, se o delineamento D tivesse sido utilizado no experimento relatado em Li *et al.* (2013), os resultados obtidos sobre o processo de separação do complexo extrato bruto de folhas de *Rosa sericea* poderiam ter agregado mais informação, uma vez que este possui, em geral, melhores capacidades preditivas.

4.2 Exemplo 2 - Comparação dos delineamentos obtidos através dos critérios D_S , DP_S e Composto

Para este exemplo comparamos os delineamentos I , II , V e VI apresentados na Tabela 2 de Trinca e Gilmour (2012), todos com 3 fatores e 18 tratamentos. O delineamento I foi obtido pelo critério D_S e o II é obtido pelo critério DP_S . Já os delineamentos V e VI são obtidos por meio do Critério Composto. A região experimental utilizada para os cálculos e elaboração dos VDGs foi a esférica, e o modelo de regressão linear utilizado foi o de segunda ordem completo, isto é, incluindo todas as interações entre os fatores. Os VDGs obtidos estão apresentados na Figura 2.

Estudando os VDGs apresentados na Figura 2 verificamos que o delineamento que apresentou maiores valores para a variância da predição é o delineamento II , tendo portanto baixo desempenho para predição. Comparando o delineamento V com o I , o delineamento V é mais instável e apresenta valores maiores para a variância da predição para uma boa parte dos raios, sendo assim não atrativos para predição. Finalmente restam os delineamentos I e o VI para comparação, e apesar de o VI ser mais instável que o I , apresenta variâncias menores até por volta do raio 0,65, e assim sendo é o mais atrativo, uma vez que fornecerá uma superfície de resposta que será capaz de realizar predições com variâncias menores, e por conseguinte margens de erros menores, para a maior parte da região experimental, o que já era de se esperar uma vez que é obtido por meio do Critério Composto. A escolha do delineamento a ser usado deve levar em conta diversas propriedades e, dificilmente, existirá um delineamento com propriedades ótimas em todos os aspectos. Neste exemplo em particular, temos que o delineamento I não permite estimação precisa do erro puro, enquanto que o VI permite, com graus de liberdade mais

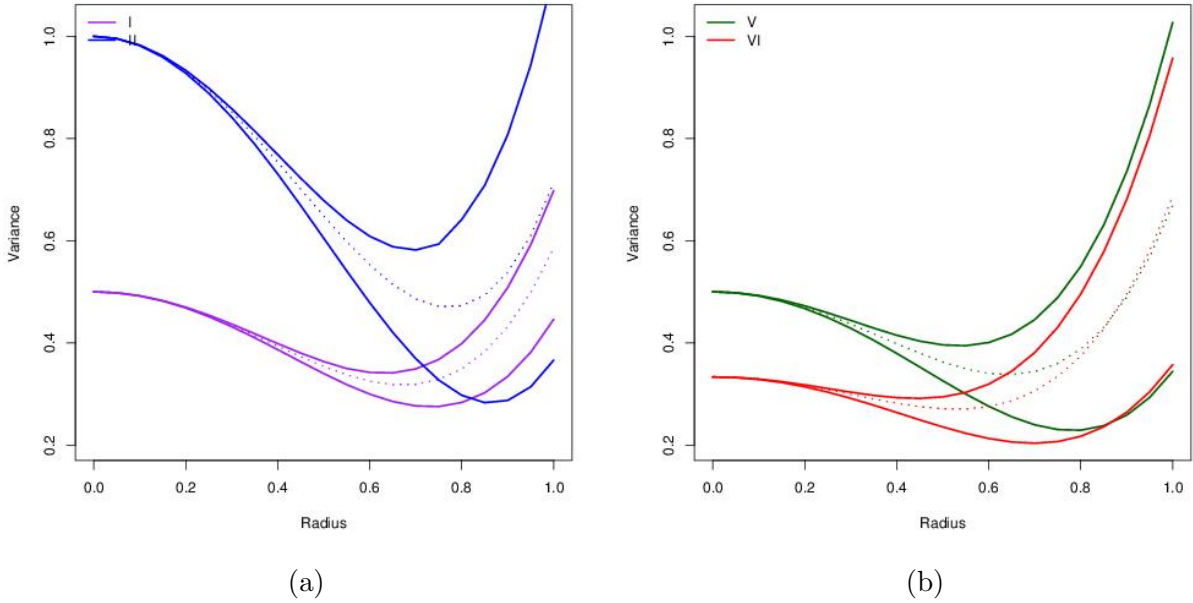


Figura 2: VDGs dos delineamentos I , II , V e VI apresentados na Tabela 2 de Trinca e Gilmour (2012).

bem distribuídos para a estimação do erro puro e do erro devido à falta de ajuste do polinômio. Estas propriedades, aliadas com o desempenho preditivo mostrado pelos VDGs, sugerem que o delineamento VI é uma boa opção de uso prático.

5 Conclusão

O objetivo tanto da implementação de um programa em linguagem C para a metodologia do VDG quanto da conexão entre este e o programa estatístico R através da instalação de um pacote, foram atingidos. Verificando de forma geral os resultados apresentados na Figura 1 e na Figura 2, concluímos que o estudo de delineamentos distintos podem ser facilmente realizados, quanto à capacidade preditiva, utilizando-se o pacote *Dispersion* construído neste trabalho.

Assim sendo, esperamos que com a futura disponibilização do pacote *Dispersion* ao CRAN, a ferramenta do VDG seja amplamente utilizada por pesquisadores que utilizam o R para que possam planejar melhor seus experimentos e, conseqüentemente, obter resultados mais informativos nas pesquisas experimentais.

Referências

- ATKINSON, A. C.; DONEV, A. N. e TOBIAS, R. D. **Optimum experimental designs, with SAS**. Oxford: Oxford University Press, 2007. 511p.
- BOX, G. E. P.; DRAPER, N. R. **Empirical model-building and response surfaces**. New York: John Wiley & Sons, 1987. 663p.
- CHAMBERS, J.; HAND, D.; HÄRDLE, W. **Software for data analysis - programming with R**. New York: Springer, 2008. 497p.
- GENTELEMAN, R. **R - programming for bioinformatics**. Washington: CRC Press, 2009. 305p.
- JENSEN, G.; MYERS, R. H. Graphical assessment of the prediction capability of response surface designs. **Technometrics**. v.31, n.2, p.159-171, 1989.
- LI, J. R.; LI, M.; XIA, B.; DING, L. S.; XU, H. X.; ZHOU, Y. Efficient optimization of ultra-high-performance supercritical fluid chromatographic separation of *Rosa sericea* by response surface methodology. **Journal of Separation Science**. v.36, p.2114-2120, 2013.
- PRESS, H. W.; TEUKOLSKY, S. A.; VETTERLING, W. T.; FLANNERY B. P. **Numerical recipes in C - the art of scientific computing**, 2^a ed. New York: Cambridge University Press, 1992. 965p.
- TRINCA, L. A.; GILMOUR, S. G. Optimum design of experiments for statistical inference (with discussion). **Journal of the Royal Statistical Society, Series C (Applied Statistics)**, v.61, p.345-401, 2012.
- VINING, G. G. Computer program for generating variance dispersion graphs. **Journal of Quality Technology**. v.25, n.1, p.45-58, 1993.

A Delineamentos comparados para o Exemplo 1

Box-Behnken					Central Composto					D				
X1	X2	X3	X4	X5	X1	X2	X3	X4	X5	X1	X2	X3	X4	X5
-1	-1	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	1	0	0	0	-1	-1	-1	-1	1	-1	-1	-1	-1	1
1	-1	0	0	0	-1	-1	-1	1	-1	-1	-1	-1	1	-1
1	1	0	0	0	-1	-1	-1	1	1	-1	-1	-1	1	1
0	0	-1	-1	0	-1	-1	1	-1	-1	-1	-1	0	-1	1
0	0	-1	1	0	-1	-1	1	-1	1	-1	-1	0	1	0
0	0	1	-1	0	-1	-1	1	1	-1	-1	-1	1	-1	-1
0	0	1	1	0	-1	-1	1	1	1	-1	-1	1	0	1
0	-1	0	0	-1	-1	1	-1	-1	-1	-1	-1	1	1	-1
0	-1	0	0	1	-1	1	-1	-1	1	-1	0	0	0	-1
0	1	0	0	-1	-1	1	-1	1	-1	-1	0	1	-1	0
0	1	0	0	1	-1	1	-1	1	1	-1	0	1	1	1
-1	0	-1	0	0	-1	1	1	1	-1	-1	-1	1	-1	-1
-1	0	1	0	0	-1	1	1	-1	1	-1	1	-1	-1	1
1	0	-1	0	0	-1	1	1	1	-1	-1	1	-1	0	1
1	0	1	0	0	-1	1	1	1	1	-1	1	-1	1	-1
0	0	0	-1	-1	1	-1	-1	-1	-1	-1	1	1	-1	-1
0	0	0	-1	1	1	-1	-1	-1	1	-1	1	1	-1	1
0	0	0	1	-1	1	-1	-1	1	-1	-1	1	1	1	-1
0	0	0	1	1	1	-1	-1	1	1	-1	1	1	1	1
0	-1	-1	0	0	1	-1	1	-1	-1	0	-1	-1	-1	-1
0	-1	1	0	0	1	-1	1	-1	1	0	-1	1	-1	1
0	1	-1	0	0	1	-1	1	1	-1	0	-1	1	1	-1
0	1	1	0	0	1	-1	1	1	1	0	0	-1	0	0
-1	0	0	-1	0	1	1	-1	-1	-1	0	0	-1	1	1
-1	0	0	1	0	1	1	-1	-1	1	0	1	0	-1	0
1	0	0	-1	0	1	1	-1	1	-1	0	1	0	1	1
1	0	0	1	0	1	1	-1	1	1	0	1	1	0	-1
0	0	-1	0	-1	1	1	1	-1	-1	1	-1	-1	-1	1
0	0	-1	0	1	1	1	1	-1	1	1	-1	-1	0	0
0	0	1	0	-1	1	1	1	1	-1	1	-1	-1	1	-1
0	0	1	0	1	1	1	1	1	1	1	-1	-1	1	1
-1	0	0	0	-1	-1	0	0	0	0	1	-1	0	0	-1
-1	0	0	0	1	1	0	0	0	0	1	-1	1	-1	-1
1	0	0	0	-1	0	-1	0	0	0	1	-1	1	-1	0
1	0	0	0	1	0	1	0	0	0	1	-1	1	1	1
0	-1	0	-1	0	0	0	-1	0	0	1	0	0	-1	1
0	-1	0	1	0	0	0	1	0	0	1	0	1	1	-1
0	1	0	-1	0	0	0	0	-1	0	1	1	-1	-1	-1
0	1	0	1	0	0	0	0	1	0	1	1	-1	-1	1
0	0	0	0	0	0	0	0	0	-1	1	1	-1	1	-1
0	0	0	0	0	0	0	0	0	1	1	1	-1	1	1
0	0	0	0	0	0	0	0	0	0	1	1	1	-1	-1
0	0	0	0	0	0	0	0	0	0	1	1	1	-1	1
0	0	0	0	0	0	0	0	0	0	1	1	1	0	1
0	0	0	0	0	0	0	0	0	0	1	1	1	1	0

B Funções do pacote *Dispersion*

```
Variance.dispersion <- function (DESIGN, REGION, DES.TYPE, GRAPH.TYPE, ORDER, NWPT, NSPT,
ETA, PRIN, RADII, WEIGHTING, SCALE, FRACTION, SEARCH, FTOL, ITMAX, NLOOPS)
{
  if (missing(REGION))
    REGION = 2
  if (missing(DES.TYPE))
    DES.TYPE = 1
  if (missing(GRAPH.TYPE))
    GRAPH.TYPE = 1
  if (missing(ORDER))
    ORDER = 2
  if (missing(NWPT))
    NWPT = 1
  if (missing(NSPT))
    NSPT = 1
  if (missing(ETA))
    ETA = 1
  if (missing(PRIN))
    PRIN = 2
  if (missing(RADII))
    RADII = 20
  if (missing(WEIGHTING))
    WEIGHTING = 2
  if (missing(SCALE))
    SCALE = 1
  if (missing(FRACTION))
    FRACTION = 1
  if (missing(SEARCH))
    SEARCH = 1
  if (missing(FTOL))
    FTOL = 1e-05
  if (missing(ITMAX))
    ITMAX = 5000
  if (missing(NLOOPS))
    NLOOPS = 10000
  MAX = matrix(1, ncol = dim(DESIGN)[2] + 2, nrow = RADII +
1)
  MIN = matrix(1, ncol = dim(DESIGN)[2] + 2, nrow = RADII +
1)
  GRAPH = matrix(1, ncol = 4, nrow = RADII + 1)
  NDPTS = dim(DESIGN)[1]
  K = dim(DESIGN)[2]
  DESIGN2 = as.double(t(DESIGN))
  WEIGHTING = WEIGHTING - 1
  REGION = REGION - 1
  NETA = 1
  MAX2 = as.double(t(MAX))
  MIN2 = as.double(t(MIN))
  GRAPH2 = as.double(t(GRAPH))
  Confi = c(SEARCH, NDPTS, K, ORDER, REGION, WEIGHTING, RADII,
DES.TYPE, GRAPH.TYPE, NWPT, NSPT, NETA, PRIN, NLOOPS,
ITMAX)
  RESULTS = (.C("principal", as.integer(Confi), as.double(FTOL),
as.double(SCALE), as.double(FRACTION), as.double(ETA),
as.double(DESIGN2), as.double(GRAPH2), as.double(MAX2),
as.double(MIN2)))
```



```

GRAPH2 = RESULTS[[7]]
MAX2 = RESULTS[[8]]
MIN2 = RESULTS[[9]]
dim(MAX2) = c((K + 2), RADII + 1)
dim(MIN2) = c((K + 2), RADII + 1)
dim(GRAPH2) = c(4, RADII + 1)
MAX = t(MAX2)
MIN = t(MIN2)
GRAPH = t(GRAPH2)
POS = c("X[0]", "X[1]", "X[2]", "X[3]", "X[4]", "X[5]", "X[6]")
POSITION = c(rep("X[0]", K))
for (I in 1:K) POSITION[I] = POS[I]
dimnames(MAX) = list(seq(0, RADII), c("RAD", "MAXIMUM", POSITION))
dimnames(MIN) = list(seq(0, RADII), c("RAD", "MINIMUM", POSITION))
dimnames(GRAPH) = list(seq(0, RADII), c("RAD", "MAXIMUM",
    "MINIMUM", "AVERAGE"))
OPTIONS = rep("NOTHING", 4)
if (DES.TYPE == 1)
    OPTIONS[1] = "completely randomized design"
if (DES.TYPE == 2)
    OPTIONS[1] = "split-plot design"
if (REGION == 0)
    OPTIONS[2] = "spherical region"
if (REGION == 1)
    OPTIONS[2] = "cuboidal region"
if (GRAPH.TYPE == 1)
    OPTIONS[3] = "VDG"
if (GRAPH.TYPE == 2)
    OPTIONS[3] = "DVDG"
if (ORDER == 1)
    OPTIONS[4] = "first-order polynomial"
if (ORDER == 2)
    OPTIONS[4] = "second-order polynomial"
if (REGION == 1)
    GRAPH = GRAPH[, -4]
return(list(MAX = MAX, MIN = MIN, GRAPH = GRAPH, OPTIONS = OPTIONS))
}

```

```

Plot.dispersion <- function (DESIGN1, DESIGN2, SCALE, LEGEND, COLOR)
{
  if (missing(SCALE)) {
    if (missing(DESIGN2)) {
      SCALE = c(min(DESIGN1[, 3]), max(DESIGN1[, 2]))
    }
    else {
      SCALE = c(min(DESIGN1[, 3], DESIGN2[, 3]), max(DESIGN1[,
        2], DESIGN2[, 2]))
    }
  }
  if (missing(COLOR)) {
    COLOR = c("dark green", "red")
  }
  if (missing(DESIGN2)) {
    PLOT2 = 0
  }
  else {
    PLOT2 = 1
  }
}

```

```

CONTROL = 1
if (missing(LEGEND))
  CONTROL = 0
REGION1 = dim(DSIGN1)[2]
if (PLOT2 == 1)
  REGION2 = dim(DSIGN2)[2]
plot(DSIGN1[, 1], DESIGN1[, 2], lwd = 2, col = COLOR[1],
     type = "l", ylim = c(SCALE[1] - 0.001, SCALE[2] + 0.001),
     xlab = "Radius", ylab = "Variance")
lines(DSIGN1[, 1], DESIGN1[, 3], lwd = 2, col = COLOR[1])
if (REGION1 == 4)
  lines(DSIGN1[, 1], DESIGN1[, 4], lwd = 2, lty = 3, col = COLOR[1])
if (PLOT2 == 1) {
  lines(DSIGN2[, 1], DESIGN2[, 2], lwd = 2, col = COLOR[2])
  lines(DSIGN2[, 1], DESIGN2[, 3], lwd = 2, col = COLOR[2])
  if (REGION2 == 4)
    lines(DSIGN2[, 1], DESIGN2[, 4], lwd = 2, lty = 3,
          col = COLOR[2])
}
if (CONTROL == 1) {
  if (PLOT2 == 1)
    legend("topleft", LEGEND, lty = c(1, 1), col = COLOR,
           lwd = c(2, 2), bty = c("n", "n"))
  if (PLOT2 == 0)
    legend("topleft", LEGEND[1], lty = c(1), col = COLOR[1],
           lwd = c(2), bty = c("n"))
}
return(SCALE = SCALE)
}

```