TEORIA DE VALORES EXTREMOS E TAMANHO AMOSTRAL PARA O MELHORAMENTO GENÉTICO DO QUANTIL MÁXIMO EM PLANTAS

José Alfredo Diaz ESCOBAR[1]
Marcos Deon Vilela de RESENDE[1]
Camila Ferreira AZEVEDO[1]
Fabyano Fonseca SILVA[2]
Márcio Henrique Pereira BARBOSA[3]
Andrei Caíque Pires NUNES[4]
Rodrigo Silva ALVES[4]
Moysés NASCIMENTO[1]

- RESUMO: Este trabalho objetivou propor e avaliar uma metodologia estatística para o melhoramento genético do valor extremo das distribuições de caracteres quantitativos. Essa abordagem baseia-se nos quantis superiores da GEV (Distribuição de Valores Extremos Generalizada) e da DEV (Distribuições de Valores Extremos) dos valores genotípicos individuais entre e dentro de famílias de plantas ou animais. Usando bases de dados reais e simulados de progênies ou famílias de cana-de-açúcar, distribuições de valores extremos (Gumbel, Fréchet e Weibull) foram ajustadas aos máximos das famílias. Simulações estocásticas e reamostragens de dados experimentais indicaram consistentemente que a avaliação de 200 famílias maximiza a eficiência do melhoramento visando à seleção de indivíduos extremos. A distribuição Weibull foi a de melhor ajuste (seguida pela Gumbel) e indicou aumento da eficiência seletiva de 1,10 (ganho de 10%) quando se passa de 20 para 100 indivíduos por família e de 1,12 (ganho de 2%) quando se passa de 100 para 200 indivíduos. Esses números são aproximadamente constantes independentemente do número de famílias avaliadas. Uma boa opção prática seria a avaliação de 200 famílias com 100 indivíduos, num total de 20000 indivíduos. A metodologia é adequada também para classificar famílias pela capacidade de geração de indivíduos superiores e informar os tamanhos amostrais em cada família para capturar esses indivíduos.

- PALAVRAS-CHAVE: Distribuições de probabilidade; propagação vegetativa; indivíduo extremo; tamanho de família; número de famílias.

[1] Universidade Federal de Viçosa - UFV, Programa de Pós-Graduação em Estatística Aplicada e Biometria, Departamento de Estatística, CEP: 36570-000, Viçosa, MG, Brasil. E-mail: *afro77777@gmail.com*; *marcos.deon@gmail.com*; *camila.azevedo@ufv.br*; *moysesnascim@ufv.br*
[2] Universidade Federal de Viçosa - UFV, Departamento de Zootecnia, CEP: 36570-000, Viçosa, MG, Brasil. E-mail: *fabyanofonseca@ufv.br*
[3] Universidade Federal de Viçosa - UFV, Departamento de Fitotecnia, CEP: 36570-000, Viçosa, MG, Brasil. E-mail: *barbosa@ufv.br*
[4] Universidade Federal de Viçosa - UFV, Programa de Pós-Graduação em Genética e Melhoramento, CEP: 36570-000, Viçosa, MG, Brasil. E-mail: *andreicaique@yahoo.com.br*; *ralves.ufla@gmail.com*

1 Introdução

Dentre os objetivos dos programas de melhoramento genético de plantas de propagação assexuada (como a cana-de-açúcar e o eucalipto) e autógamas encontra-se o de selecionar indivíduos extremos ou segregantes transgressivos. Assim, é conveniente encontrar famílias com distribuições de caudas mais longas ou mesmo assimétricas, já que elas têm uma maior tendência de gerar indivíduos excepcionais. Os métodos de seleção comumente utilizados no melhoramento dessas espécies enquadram-se na classe BLUP sob os conceitos de média aritmética e média harmônica (RESENDE e BARBOSA, 2005), os quais não levam em consideração a ocorrência de valores extremos dentro das famílias. Diante do exposto, e seguindo a sugestão de Resende e Barbosa (2005), este trabalho teve como objetivo propor e avaliar uma metodologia estatística para o melhoramento do máximo ou valor extremo das distribuições e não necessariamente das médias das distribuições. Essa abordagem baseia-se nos quantis superiores da GEV (Distribuição de Valores Extremos Generalizada) dos BLUP's (melhores preditores lineares não viesados) genotípicos individuais entre e dentro de famílias, como forma de prever o aumento da ocorrência de valores extremos em função do aumento do tamanho da família (seleção de indivíduos extremos dentro de família) e também do número de famílias utilizado para representar uma população (seleção de indivíduos extremos em toda a população).

Os valores extremos, por definição, são poucos, e suas estimações frequentemente são feitas para níveis de um processo que são muito maiores que os observados. Deste modo, o objetivo essencial da teoria dos valores extremos é a extrapolação da informação. Uma vez que não se tem fundamentos empíricos ou físicos para desenvolver uma regra de extrapolação, a teoria assintótica é utilizada para encontrar as distribuições limites dos valores extremos. A teoria de valores extremos (TVE) fornece uma base sólida e uma estrutura para a extrapolação, levando a estimadores naturais para as quantidades correspondentes, como são os quantis extremos. Assim, a TVE é reconhecida como uma disciplina única na estatística porque gera técnicas e modelos para descrever o inusitado (raro) ao invés do habitual (COLES, 2001).

As ideias principais da teoria de valores extremos podem ser descritas da seguinte maneira. Dado uma amostra de variáveis aleatórias independentes e identicamente distribuídas, pretende-se analisar o comportamento de qualquer uma das estatísticas de ordem, como o mínimo, o máximo, ou o limiar de um nível crítico, onde as estatísticas de ordem devem ter funções de distribuição *max-estável*, já que por definição as distribuições do máximo e do mínimo são degeneradas. Portanto, deve-se usar um método que permita aproveitar o argumento assintótico, como o teorema dos Tipos Extremos, para aproximar a distribuição da estatística de ordem por uma dentre as três famílias de distribuições interessantes a este propósito, das quais a Gumbel (Tipo I), Fréchet (Tipo II), ou Weibull (Tipo III), que pertencem à classe DEV e que são casos especiais da GEV.

O modelo estatístico utilizado para a predição de indivíduos superiores ou extremos baseia-se na Distribuição de Valores Extremos Generalizada (GEV) proposto por von Mises (1936) e Jenkinson (1955), dada por:

$$G(x\backslash\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, x \in \mathbb{R}.$$

Definido no conjunto $\left\{x: 1+\xi\left(\frac{x-\mu}{\sigma}\right)>0\right\}$, os parâmetros μ, σ e ξ são chamados de localização, escala e forma respectivamente, e os valores que podem tomar são $-\infty \leq \mu \leq \infty$, $\sigma > 0$, e $-\infty \leq \xi \leq \infty$. Se $\xi \to 0$ se diz que a distribuição limite é Gumbel (tipo *I*), se $\xi > 0$ tem-se a distribuição Fréchet (tipo *II*), e se $\xi < 0$ a distribuição obtida é Weibull (tipo *III*) (COLES, 2001). As famílias das distribuições anteriores são conhecidas como Distribuições de Valores Extremos (DEV), as quais estão definidas pelo teorema de Fisher –Tippet e Gnedenko (CASTILLO *et al*., 2005; BOVIER, 2010; COLES, 2001) como:

Distribuição Gumbel: $G(x) = \exp\left\{-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}, -\infty < x < \infty$

Distribuição Fréchet: $G(x) = \begin{cases} 0, & se\ x \leq b \\ \exp\left\{-\left(\frac{x-b}{a}\right)^{-\alpha}\right\}, & se\ x > b \end{cases}$

Distribuição Weibull: $G(x) = \begin{cases} exp\left\{-\left[-\left(\frac{x-b}{a}\right)^{-\alpha}\right]\right\}, & se\ x \leq b \\ 1, & se\ x > b \end{cases}$

em que os parâmetros $a > 0$ e $b \in R$, e no caso das famílias Fréchet e Weibull $\alpha > 0$.

Os três tipos de famílias de distribuições são combinadas numa única família de distribuições com parametrização comum, sendo esta formulação chamada GEV. A GEV tem três parâmetros: μ, parâmetro de localização, σ, parâmetro de escala, e ξ, parâmetro de forma. Utilizar a GEV diminui muito o esforço estatístico e computacional, já que se pode realizar uma inferência de ξ, ou seja, a amostra escolhida determinará qual das famílias é a mais adequada para realizar as análises, sem necessidade de realizar suposições sobre o tipo de DEV que se deve adotar (COLES, 2001). Outra abordagem para estudar a predição de valores extremos refere-se ao uso da técnica de regressão quantílica (CAI, 2013; SCHAUMBURG, 2012; CHERNOZHUKOV e FERNÁNDEZ-VAL, 2011).

É importante destacar que um dos campos com maiores aplicações da teoria do valor extremo é no planejamento de estruturas, que devem resistir a algum fenômeno ambiental como o nível do mar, velocidade do vento, nível de um rio ou represa, concentração de contaminantes, chuvas e ondas. Se o fenômeno é muito intenso, a estrutura falhará e desse modo, é necessário projetar esta de maneira que a probabilidade de falha em função do evento natural extremo seja menor (COLES, 2001).

Para essas previsões, emprega-se o período de retorno associado à ocorrência de um evento raro (nível de retorno) típico da distribuição ajustada. No caso do melhoramento vegetal, o período de retorno pode ser caracterizado como o tamanho amostral necessário para a ocorrência do genótipo excepcional. Dessa forma, com o uso de uma distribuição de valor extremo para predição do máximo das distribuições dos indivíduos, é possível prever o comportamento da eficiência seletiva para os máximos associados a vários tamanhos de famílias e de populações experimentais. Apesar da grande relevância da TVE, estudos que reportam sua aplicação no melhoramento de plantas são inexistentes.

Diante do exposto, este estudo teve como objetivo propor e avaliar o uso da teoria de valores extremos no melhoramento de plantas de propagação assexuada. Essa abordagem baseia-se nas distribuições dos quantis superiores, como forma de prever o aumento da

ocorrência de valores extremos em função do aumento do tamanho da família (seleção de indivíduos extremos dentro de família), do número de famílias utilizado para representar uma população (seleção de indivíduos extremos em toda a população), permitindo assim, a classificação das famílias pela capacidade de geração de indivíduos excepcionais.

Para isso, os objetivos específicos do artigo são: comparar o comportamento das diferentes classes de distribuições de probabilidade de valores extremos como ferramenta para inferência sobre o tamanho amostral necessário para a ocorrência de indivíduos extremos em programas de melhoramento genético de plantas; predizer o valor genotípico dos futuros indivíduos extremos em novas amostragens dentro de famílias e dentro de populações formadas por várias famílias; estipular os tamanhos amostrais necessários, em termos de número de indivíduos por famílias e número de famílias necessários para a ocorrência de indivíduos extremos; aferir uma metodologia para classificar as famílias ou progênies pela capacidade de geração de indivíduos superiores ou excepcionais e informar os tamanhos amostrais a serem praticados em cada família para capturar esses indivíduos.

2 Material e métodos

2.1 Metodologia estatística

O modelo estatístico utilizado para a predição de indivíduos superiores foi a Distribuição de Valores Extremos Generalizada (GEV), conforme descrita na Introdução. A GEV e as três distribuições de valores extremos foram aplicadas no seguinte contexto. Uma distribuição de valor extremo é empregada para a predição do máximo das distribuições dos indivíduos dentro de famílias e de populações experimentais. Para essas previsões, emprega-se o período de retorno associado à ocorrência de um nível do evento raro típico da distribuição ajustada. No caso, o período de retorno é interpretado como o tamanho amostral necessário para a ocorrência de determinado nível de retorno do evento raro (valor extremo com sua magnitude).

A metodologia para identificar uma observação ou valor extremo foi o Método do Bloco Máximo, que em sua forma básica consiste em dividir a amostra de tamanho m, em b blocos (famílias) de similar tamanho n (n suficientemente grande), e logo obter os valores máximos de cada família (b), assim com a nova amostra $M_{n,1}, M_{n,2}, \ldots., M_{n,b}$, foram estimados μ, σ, e ξ mediante o ajuste da GEV. Para estimar os parâmetros das amostras dos máximos utilizaram-se os métodos de Máxima Verossimilhança (MML) e bayesianos, quando se utilizou o MML foi necessário analisar o intervalo de confiança de ξ e o teste de razão de verossimilhança, conforme sugerido por Gumbel (2004), para determinar qual das famílias de distribuições era a adequada para estudar a amostra m escolhida das respectivas variáveis MMC e B.

O MML é o mais usado por sua generalidade (propriedades assintóticas dos estimadores) e flexibilidade (COLES, 2001), no entanto, na maior parte dos programas de melhoramento de plantas perenes como cana de açúcar, trabalham com conjuntos de dados pequenos, e os métodos Bayesianos proporcionam resultados precisos quando tem-se pequenas amostras, já que proporcionam uma distribuição a posteriori fidedigna para obter as inferências pertinentes, seja para grandes ou pequenas amostras (RESENDE, 2000).

Com a metodologia Bayesiana para as análises dos parâmetros das variáveis MMC e B foram usadas como a priori a distribuição normal trivariada ($\mu = 2,\ \sigma = 0{,}3,\ \xi = 0{,}04$)

e ($\mu = 15,\ \sigma = 2,\ \xi = 0{,}05$) respetivamente, e utilizaram-se métodos de simulação de Monte Carlo via Cadeias de Markov (MCMC); para o MML também foi necessário o uso de métodos numéricos para obter as estimativas dos parâmetros porque analiticamente é impossível obtê-los, entretanto, deve-se ter cautela quando se trabalha com o MML, já que as condições de regularidade necessárias para que as propriedades assintóticas associadas com os estimadores de máxima verossimilhança em algumas circunstâncias não são satisfeitas (COLES, 2001).

A função de densidade preditiva utilizada no método bayesiano para a obtenção dos parâmetros e níveis de retorno foi a seguinte:

$$Pr\{\mathrm{Z} \leq z|x_1, \ldots\ldots, x_n\} = \int_{\Theta} Pr\{\mathrm{Z} \leq z|\theta\} f(\theta|x)\, d\theta$$

em que $Pr\{\mathrm{Z} \leq z|\theta\}$ é a GEV avaliada em z, $\theta = \mu, \sigma$ e ξ e $f(\theta|x)$ é a distribuição a posteriori (CASTILLO *et al*., 2005; DE HAAN e FERREIRA, 2006; STEPHENSON e RIBATET, 2014).

Os intervalos de confiança para as estimativas dos parâmetros foram obtidos pelo método do Perfil de Verossimilhança, e na metodologia Bayesiana utilizou-se o Bootstrap Paramétrico, que emprega o procedimento do cálculo do porcentual a partir da simulação da amostra (percentis do MCMC). Para fazer qualquer tipo de predição (extrapolação), foi necessário constatar que o ajuste realizado do modelo GEV era válido, para isto utilizaram se ferramentas gráficas como: quantil-quantil (QQ), probabilidade (PP), densidade dos dados vs. densidade do modelo, ademais realizaram se os testes de aderência de Anderson Darling e Kolmogorov-Smirnov.

Após conhecer a distribuição mais adequada para analisar as variáveis em estudo e as estimativas dos parâmetros μ, σ, e ξ, obtiveram se os níveis de retornos x_p (quantis) associados aos quatro (20, 30, 43 e 63) períodos de retorno $\left(\frac{1}{p}\right)$ desejados, por meio da seguinte expressão:

$$\hat{x}_p = \begin{cases} \hat{\mu} - \dfrac{\hat{\sigma}}{\hat{\xi}}\left(1 - {y_p}^{-\hat{\xi}}\right), & para\ \xi \neq 0 \\ \hat{\mu} - \hat{\sigma} \log y_p\,, & para\ \xi = 0, \end{cases}$$

em que $y_p = -\log\ (1 - p)$.

Construíram-se os respectivos gráficos dos níveis de retorno vs. períodos de retorno, já que são muito úteis porque facilitam distinguir os resultados da extrapolação de níveis de retorno em períodos de retorno longos e também permite apresentar e validar a distribuição pertinente para a análise da amostra dos máximos, quando utiliza-se o método de Máxima Verossimilhança para estimar os parâmetros. Foram obtidos os intervalos de confiança dos níveis de retornos com três metodologias: Perfil de Verossimilhança (Ve. P), Método Delta (Aproximação Normal) e Bootstrap Paramétrico (percentis do MCMC). Essas metodologias são descritas a seguir.

2.1.1 Perfil de verossimilhança

O método do Perfil de Verossimilhança consiste em maximizar a função de verossimilhança com base em parâmetros definidos previamente (SPROTT, 2000; CASTILLO *et al*., 2005), ou seja:

$$L_p(\theta^{(1)}) = \max_{\theta^{(2)}|\theta^{(1)}} \mathrm{L}(\theta^{(1)}, \theta^{(2)}).$$

A avaliação numérica do perfil da verossimilhança para qualquer um dos parâmetros μ, σ, ξ ou os níveis de retorno $\hat{x}_p$ (para $\hat{x}_p$ precisa-se de uma nova parametrização do modelo GEV) é obtida de uma forma relativamente prática, por exemplo, se quisermos obter o perfil da verossimilhança de ξ, fixa-se ξ= ξo e maximizamos a log-verossimilhança com relação aos parâmetros restantes. Logo se reproduz isso para um determinado número de valores de ξ=ξo, os respectivos valores maximizados da log-verossimilhança compõem o perfil da log-verossimilhança para ξ, com o procedimento geral para construção de intervalos de confiança, podem-se obter os respectivos intervalos aproximados para cada parâmetro (COLES, 2001; CASTILLO *et al*., 2005).

2.1.2 Método Delta

O método Delta admite a normalidade aproximada (convergência em distribuição) de $\hat{x}_p$ para ser utilizada na obtenção dos intervalos de confiança de x_p, portanto o método permite aproximar a $Var(\hat{x}_p)$ por meio da seguinte expressão: $Var(\hat{x}_p) \approx \nabla x_p^t \, \mathbf{V} \, \nabla x_p$ em que $\mathbf{V} = \mathbf{\Sigma}$ é a matriz de covariâncias de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ e avaliadas em $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ e ∇x_p^t definida por

$$\nabla x_p^t = \left(\frac{\partial x_p}{\partial \xi}, \frac{\partial x_p}{\partial \mu}, \frac{\partial x_p}{\partial \sigma}\right) = \left(\frac{\sigma}{\xi^2}\left(1 - y_p{}^{-\xi}\right) - \frac{\sigma}{\xi}\, y_p{}^{-\xi} \log y_p, \qquad 1, -\frac{1}{\xi}\left(1 - y_p{}^{-\xi}\right)\right).$$

2.1.3 Bootstrap paramétrico

O método de bootstrap paramétrico pressupõe que existe uma distribuição de probabilidade que originou $x_1, x_2 \ldots\ldots x_n$ (amostra aleatória de tamanho n, chamada amostra original), portanto utiliza o modelo de probabilidade pressuposto e as estimativas dos parâmetros obtidas com a amostra original, para gerar um grande número de amostras independentes $x_1^*, x_2^* \ldots\ldots x_B^*$, chamadas amostras bootstrap. θ representa o parâmetro de interesse, e uma réplica bootstrap $\theta_b^*, b = 1,2, \ldots\ldots B$, corresponde ao valor do estimador de máxima verossimilhança $\hat{\theta}$ avaliado em cada uma das B amostras bootstrap.

Após obter as B amostras bootstrap, também é admissível construir uma distribuição bootstrap para o estimador de máxima verossimilhança $\hat{\theta}$, portanto, a distribuição obtida pode ser usada para realizar inferências sobre o parâmetro em estudo. Os intervalos de confiança bootstrap percentil são obtidos pelos respectivos percentis $\frac{\alpha}{2} - ésimo$ e $(1 - \frac{\alpha}{2}) - ésimo$ da distribuição de $\hat{\theta}^*$, representada por $\hat{F}$ (EFRON e TIBSHIRANI, 1983).

As metodologias apresentadas foram comparadas por meio dos intervalos de confiança e da Eficiência Média (E_M; indicador de viés), em que 100% indica ausência de viés:

$$E_M = \frac{\bar{y}_n}{V.\ Est.}$$

em que $\bar{y}_n$ é a média dos três ou cinco melhores indivíduos, e $V.\ Est.$ é a estimativa do nível de retorno $\hat{x}_p$.

2.2 Dados simulados

Para estudar o tamanho amostral visando à maximização da eficiência do melhoramento do máximo de uma distribuição foram consideradas duas situações: simulação de dados e *um conjunto de dados reais de cana-de-açúcar*.

Foram simulados nove cenários com diferentes médias, desvios padrões e valores máximos, conforme a Tabela 1.

Tabela 1 - Cenários (C) de famílias e indivíduos simulados

	Número de indivíduos		
Número de famílias	20	100	200
20	C1	C2	C3
100	C4	C5	C6
200	C7	C8	C9

Cada cenário foi simulado cem vezes e os indivíduos nas famílias correspondem a uma distribuição normal truncada para valores positivos com tamanhos amostrais de 20, 100 e 200 indivíduos. Para os cenários C1, C2, e C3 as médias utilizadas foram 5, 10, 15 e 20 toneladas de açúcar por hectare (TPH), os desvios padrões usados para cada uma das médias anteriores foram 1, 2, 3, 4, e 5 TPH. Nos cenários C4, C5, e C6 as médias iniciam em 1,6 e incrementam-se em 1 (uma) unidade até chegar a 20,6 TPH, os desvios padrões são os mesmos dos cenários anteriores. Em referência aos cenários C7, C8, e C9 os valores das médias de TPH iniciaram com 0,3 e foram aumentando 0,3 até 12,0, também se utilizaram os desvios padrões 1, 2, 3, 4, e 5 TPH para cada uma das médias simuladas anteriormente.

Um indivíduo simulado representa a produção de uma parcela experimental. Após simular os cenários C1 até C9, foram encontrados os máximos, médias e desvios padrões das amostras dos máximos. Para encontrar o modelo adequado para analisar as amostras dos máximos, utilizou-se a metodologia GEV e para os intervalos de confiança utilizou-se o método do Perfil de Verossimilhança.

Realizaram-se os testes de Anderson Darling e Kolmogorov-Smirnov, para verificar a qualidade dos ajustes dos modelos encontrados (Gumbel, Weibul e Frechét). Após conhecer a distribuição adequada para as análises das amostras, obtiveram-se os níveis de retornos (valor do máximo) associados aos períodos de retorno (números de famílias a serem avaliadas para a obtenção do máximo) 20, 50, 100, 200, 500 e 1000. Deve ser ressaltado

que os tamanhos totais das "novas populações" variaram de 400 a 200000 indivíduos. Dessa forma, as eficiências em nível individual, de família e populacional foram determinadas.

O cálculo das eficiências para o melhoramento em cana de açúcar foi feito com relação ao aumento de indivíduos, famílias e populações. Desse modo foram utilizadas as seguintes expressões:

$$E_{ind} = \frac{Est_{F.n}}{Est_{F.20}} \; ; \; E_{fam} = \frac{Est_{F.n}}{Est_{20.n}} \; ; \; E_{pop} = \frac{Est_{F.n}}{Est_{20.20}}$$

em que:
E_{ind} é a eficiência com relação ao aumento do número de indivíduos,
$Est_{F.n}$ é a estimativa do valor máximo com F novas famílias e *n* indivíduos (F=20,50,100,200,500 e 1000; *n* = 20,100 e 200),
$Est_{F.20}$ é a estimativa do valor máximo com F novas famílias e 20 indivíduos (F=20,50,100,200,500 e 1000),
E_{fam} é a eficiência com relação ao aumento do número de famílias,
$Est_{20.n}$ é a estimativa do valor máximo com 20 novas famílias e *n* indivíduos (*n* = 20,100 e 200),
E_{pop} é a eficiência com relação ao aumento da população,
$Est_{20.20}$ é a estimativa do valor máximo com 20 novas famílias e 20 indivíduos.

2.3 Reamostragem em dados reais de cana-de-açúcar

Utilizando dados reais de cana-de-açúcar obtidos no programa de melhoramento conduzido pela UFV em Oratórios – MG, as variáveis analisadas foram massa média de colmos (MMC em kg) e teor de Brix (B em %). Foram tomados subgrupos (por meio de reamostragens) criando-se quatro subpopulações com diferentes médias, desvios padrões e valores máximos. As subpopulações foram compostas por 20, 30, 43 e 63 famílias. Das 63 famílias foram usadas 20 famílias para ajustar o modelo GEV e obter as estimativas dos níveis de retorno associados a 20, 30 e 43 "novas famílias". Das 43 "novas famílias" restantes foram escolhidas ao acaso 20 e 30 famílias, para validação das estimativas obtidas anteriormente.

Deve ser ressaltado que o processo descrito anteriormente (validação cruzada), tem como finalidade analisar a idoneidade da extrapolação dos modelos GEV na prática biométrica.

2.4 Recursos computacionais usados

Os processos de simulação e com dados reais foram realizados com o *software* R (R CORE TEAM, 2017), as bibliotecas utilizadas foram *extRemes* (GILLELAND e KATZ, 2011) o qual tem um conjunto de funções para a realização de análises dos valores extremos de um processo de interesse, utilizando o método do bloco máximo ou excessos ao longo de um limiar elevado; *in2extRemes* (GILLELAND e KATZ, 2011) proporciona um conjunto de janelas (interfaces gráficas) que resumem algumas das principais funções do pacote *extRemes*. O pacote *evd* (STEPHENSON, 2002) estende funções de simulação, distribuição, quantis e densidades para as distribuições de valores extremos paramétricos

uni e multivariadas, também fornece funções de ajuste que calculam estimativa de máxima verossimilhança para os métodos de bloco máximo e limiar uni e bivariadas; para corroborar os resultados obtidos com os métodos bayesianos do pacote *extRemes*, utilizou-se o pacote *evdbayes* (STEPHENSON e RIBATET, 2014) que fornece funções para a análise bayesiana de modelos de valores extremos, usando métodos MCMC.

O pacote *truncnorm* (TRAUTMANN *et al*., 2014) proporciona funções para encontrar as diferentes características (densidade, função de distribuição, função quantil, geração aleatória e função do valor esperado) da distribuição normal truncada.

3 Resultados e discussões

3.1 Dados simulados

O modelo mais adequado para ajustar amostras de 20 famílias independentemente do número de indivíduos é a Gumbel, sendo que o modelo Weibull também se ajusta razoavelmente para este tipo de amostras. Os níveis de retornos ($\hat{x}_p$) obtidos pelo modelo Gumbel são maiores que os conseguidos pelo modelo Weibull, situação que se reflete ao analisar os respectivos intervalos de confiança (dados não mostrados).

As estimativas obtidas de $\hat{x}_p$ para amostras de 20 famílias obtidas por meio dos modelos Weibull, sempre ficam muito perto do limite inferior do intervalo de confiança, circunstância que também é frequente para os modelos Weibull ajustados para amostras de 100 e 200 famílias (dados não mostrados). Deve-se destacar que nos três cenários em média 60% dos modelos Weibull ajustados com amostras de 20 famílias, o valor estimado do parâmetro ξ é maior que 0,5 (análise individual dos resultados das simulações). Portanto, as inferências realizadas são verossímeis, e os modelos que tiveram uma estimativa de ξ diferente da apresentada anteriormente, obtiveram estimativas pontuais e intervalos de $\hat{x}_p$ similares.

Em geral, para as amostras de 100 e 200 famílias independentemente do número de indivíduos (20, 100 e 200) por família, o modelo que melhor se ajusta aos máximos das amostras estudadas foi a Weibull. Para todos os modelos Weibull a estimativa do parâmetro ξ é maior que – 0,5 (análise individual de cada uma das 100 simulações feitas), portanto pode-se ter uma alta confiança nas estimativas pontuais e intervalos de confiança de $\hat{x}_p$ TPH (SMITH, 1985).

Os resultados preditivos dos modelos Gumbel e Weibull apresentam estabilidade e consistência nas estimativas dos períodos de retorno e seus respectivos intervalos de confiança (dados não mostrados). Na Tabela 2 encontram-se os prognósticos dos valores máximos das simulações.

Quanto ao tamanho da população total, para obtenção de eficiência de 1,42 torna-se necessária a avaliação de 50000 indivíduos, número esse que pode ser proibitivo em algumas espécies. Uma boa opção prática seria a avaliação de 200 famílias com 100 indivíduos, perfazendo um total de 20000 indivíduos, número esse corriqueiro no melhoramento de eucalipto, por exemplo. Nesse cenário, a eficiência é de 1,36. Dentre os cenários, a eficiência máxima é de 1,46 e ocorre avaliando-se 1000 famílias com 100 ou 200 indivíduos. Os valores de 1,36 e 1,42 são mais atrativos dado o esforço experimental necessário. O número adequado de família em torno de 200 foi também recomendado em estudo teórico (via avaliação numérica e simulação determinística) reportado por Resende

(1995). Além disso, Silva *et al.* (2015), em estudo com dados reais de cana-de-açúcar, reportaram um número ideal de 100 genótipos a serem avaliados dentro de famílias superiores desta espécie corroborando com os resultados simulados apresentados no presente trabalho.

Tabela 2 - Valores do *máximo* (obtidos pela distribuição Weibull) para populações compostas por n_{fam} famílias representadas por n_{ind} indivíduos e eficiências do aumento de n_{fam} (E_{fam}), nind (E_{ind}) e população total (E_{pop}) para o caráter produtividade de açúcar em tonelada por hectare (TPH)

n_{fam}	n_{ind}	n_{pop}	$Máximo$	E_{ind}	E_{fam}	E_{pop}	Média de E_{fam}*	Média de E_{ind}**
20	20	400	20,03	1,00	1,00	1,00		1,00
20	100^{+}	2000	22,43	1,12	1,00	1,12		$1,10^{+}$
20	200	4000	23,37	1,17	1,00	1,17	1,00	1,12
50	20	1000	22,17	1,00	1,11	1,11		
50	100	5000	24,62	1,11	1,10	1,23		
50	200	10000	25,38	1,14	1,09	1,27	1,10	
100	20	2000	23,53	1,00	1,17	1,17		
100	100	10000	25,99	1,10	1,16	1,30		
100	200	20000	26,6	1,13	1,14	1,33	1,16	
200	20	4000	24,73	1,00	1,23	1,23		
200^{++}	100^{++}	20000^{++}	27,18	1,10	1,21	$1,36^{++}$		
200	200	40000	27,61	1,12	1,18	1,38	1,21	
500	20	10000	26,08	1,00	1,30	1,30		
500	100	50000^{+++}	28,49	1,09	$1,27^{+++}$	$1,42^{+++}$		
500^{++++}	200	100000	28,69	1,10	1,23	1,43	$1,27^{++++}$	
1000	20	20000	26,96	1,00	1,35	1,35		
1000	100	100000	29,32	1,09	1,31	1,46		
1000	200	200000	29,34	1,09	1,26	1,46	1,31	

$^{+}$ Ótimo para número de indivíduos; $^{++}$ Ótimo para situação prática; $^{+++}$ Ótimo para tamanho de população; $^{++++}$ Ótimo para número de famílias; * Média da eficiência por meio dos diferentes tamanhos de família;** Média da eficiência por meio dos diferentes números de famílias; Média geral do caráter nos vários cenários igual a 6,4 com desvio padrão de 5,5

O aumento da eficiência seletiva com o aumento do tamanho de família é em torno de 1,10 quando se passa de 20 para 100 indivíduos por família e de 1,12 quando se passa de 100 para 200 indivíduos. Esses números são aproximadamente constantes independentemente do número de famílias avaliadas. Assim a inferência sobre o tamanho adequado de família pode ser realizada de forma independente ao número de famílias avaliadas e os resultados remetem ao uso de 100 indivíduos por família, não sendo compensatório duplicar esse número, pois o acréscimo seria da ordem de apenas 2%.

O comportamento da eficiência seletiva em função do tamanho da população experimental total é apresentado na Figura 1. A distribuição dos pontos revela um comportamento assintótico da eficiência com grande aproximação à assíntota em 20000 indivíduos.

É importante destacar que nas análises feitas em todos os cenários estudados, nenhuma das amostras simuladas indicou a distribuição Fréchet, como a mais adequada das distribuições para realizar as respetivas análises dos valores extremos. Entretanto, recomenda-se realizar outras pesquisas com a análise de valores extremos de outras variáveis biométricas em diferentes cenários, de tal maneira que pudesse aferir sobre a aplicabilidade das famílias de distribuições Fréchet nas análises dos valores extremos em variáveis utilizadas no melhoramento genético de plantas.

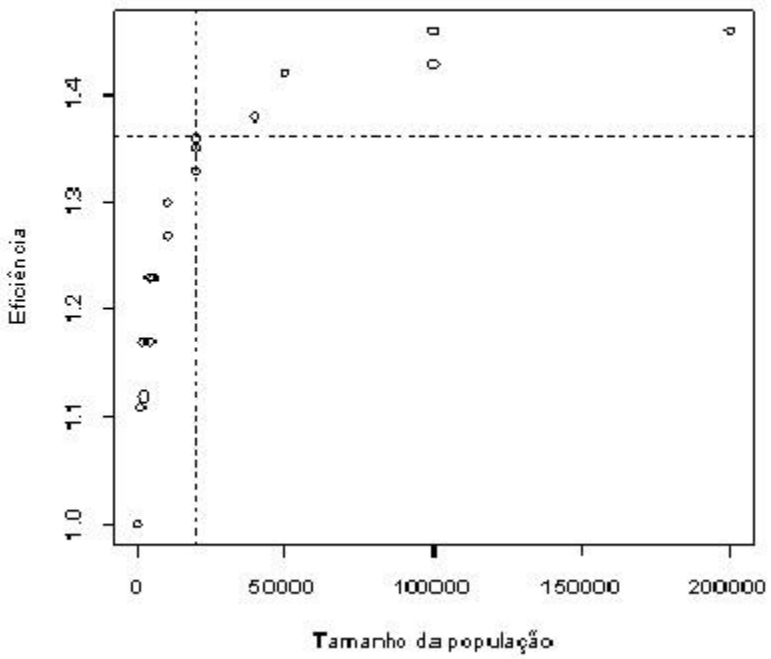

Figura 1 - Comportamento da eficiência seletiva em função do tamanho da população experimental total.

3.2 Dados reais

De acordo com o teste da razão de verossimilhança para os modelos ajustados GEV ($\xi = 0$ e $\xi \neq 0$) para a variável MMC o modelo adequado para analisá-la é a Gumbel p-valor = 0,758 (associado à hipótese $\xi = 0$), situação que corrobora com o respectivo intervalo de confiança de ξ (-0,232 ; 0,290). Para a variável Brix, o teste apresenta um p-valor = 0,014 (associado à hipótese $\xi \neq 0$), indicando que a distribuição mais adequada

para as análises é diferente da Gumbel. Dessa forma, analisou-se o intervalo de confiança correspondente (-0,636; -0,152), determinando que a distribuição mais adequada é a Weibull (Tabela 3).

Deve-se ter cautela com as estimativas do modelo Weibull, dado que ξ (-0,593) está entre -1 e -0,5, advertindo que estes não possuem as propriedades assintóticas (consistência, eficiência, invariância e normalidade) dos estimadores (SMITH, 1985). Contudo, as estimativas obtidas para os parâmetros μ, σ, ξ por meio dos métodos bayesianos e o MML, assim como seus respectivos intervalos de confiança dos modelos ajustados para as variáveis MMC e Brix, apresentam resultados similares (Tabela 3). Essa situação reflete flexibilidade da teoria dos valores extremos para ser aplicada no melhoramento genético de plantas.

Os diagnósticos dos modelos Gumbel e Weibull ajustados com o MML para as amostras dos máximos de MMC e Brix, evidenciam que estes ajustam-se razoavelmente às amostras dos máximos estudadas (dados não mostrados), sendo este resultado confirmado pelos testes de Kolmogorov-Smirnov (p-valor = 0,463 e p-valor = 0,659) e Anderson-Darling (p-valor = 0,062 e p-valor = 0,608). Dessa maneira, é possível analisar os níveis de retorno desejados.

Os modelos Gumbel e Weibull ajustados com os métodos bayesianos, também podem ser considerados admissíveis para realizar as inferências correspondentes (dados não mostrados). É importante mencionar que as taxas de aceitação dos parâmetros nos modelos estudados ficaram entre 0,399 e 0,632, intervalo que é razoável para realizar as respectivas analises bayesianas (BESAG, 2001).

Tabela 3 - Estimativas e intervalos de confiança dos parâmetros μ, σ, ξ dos modelos ajustados para variáveis massa média de colmos (MMC) e teor de Brix (B)

Variável	Método	Estimador	Estimativa	Intervalos
MMC	MLE	$\hat{\mu}$	2,175	(2,027; 2,332)
		$\hat{\sigma}$	0,316	(0,247; 0,455)
		$\hat{\varepsilon}$	-0,042	(-0,232; 0,29)$^{+}$
	Bayesiana	$\hat{\mu}$	2,179	(1,979; 2,369)
		$\hat{\sigma}$	0,341	(0,242; 0,522)
		$\hat{\varepsilon}$	-0,044	(-0,127; 0,042)
Brix	MLE	$\hat{\mu}$	15,612	(15,374; 16,607)
		$\hat{\sigma}$	2,112	(1,970; 3,612)
		$\hat{\varepsilon}$	-0,593	(-0,636; -0,152)*
	Bayesiana	$\hat{\mu}$	15,642	(14,414; 16,564)
		$\hat{\sigma}$	2,304	(1,627; 3,287)
		$\hat{\varepsilon}$	-0,587	(-0,704; -0,477)

MMC: massa média de colmos; Brix; MLE: Máxima Verossimilhança; $^{+}$Gumbel; *Weibull.

Pode-se verificar que para 20 novas famílias as estimativas pontuais das toneladas máximas de MMC ($\hat{x}_p$) obtidas pelo método bayesiano (MB) e o MML, são maiores que o maior indivíduo dessas famílias, sendo os respectivos intervalos de confiança e HPD condizentes com este resultado (Tabela 4).

Com 30 e 43 novas famílias as estimativas pontuais de ambos os métodos são menores que o maior indivíduo estudado nessas amostras, não obstante, o intervalo de confiança obtido pela metodologia do perfil de verossimilhança na amostra de 43 famílias abrange o valor do maior indivíduo. Deve ser ressaltado que em ambas amostras os intervalos HPD incluem os valores analisados (Tabela 4).

Tabela 4 - Comparação das estimativas pontuais e intervalos de confiança dos níveis de retorno com os valores máximos das amostras, e resultados das acurácias das médias dos três e cinco melhores indivíduos amostrais da variável MMC

Nov. Fam.	Máx	$\overline{y_3}$	$\overline{y_5}$	Métodos	Estimativa do máximo para o caráter	Intervalos	Eficiência Média	
							$\overline{y_3}$	$\overline{y_5}$
n = 20	3,04	2,85	2,77	Ve. P	3,114	(2,894; 3,599)	91,5%	89,1%
				A.normal	3,114	(2,732; 3,496)	91,5%	89,1%
				Q.MCMC	3,349	(2,825; 4,710)	85,0%	83,0%
n = 30	3,93	3,34	3,12	Ve. P	3,244	(2,900; 3,784)	102,9%	96,2%
				A.normal	3,244	(2,822; 3,667)	102,9%	96,2%
				Q.MCMC	3,546	(2,916; 5,358)	94,1%	88,0%
n = 43	3,93	3,34	3,12	Ve. P	3,36	(3,053; 3,952)	99,3%	92,9%
				A.normal	3,36	(2,901; 3,819)	99,3%	92,9%
				Q.MCMC	3,732	(2,987; 5,978)	89,4%	83,6%
n = 63	3,93	3,48	3,30	Ve. P	3,482	(3,105; 4,122)	100,1%	94,7%
				A.normal	3,482	(2,984; 3,979)	100,1%	94,7%
				Q.MCMC	3,943	(3,058; 6,806)	88,4%	83,6%

Nov. Fam.: novas famílias - Máx.: Valor máximo - $\bar{y}_3$:Média dos três melhores indivíduos- $\bar{y}_5$:Média dos cinco melhores indivíduos - Ve. P: Perfil de Verossimilhança - A. Normal: Aproximação Normal – Q.MCMC: Quantis de Monte Carlo via Cadeias de Markov.

Quando avaliou se as 63 famílias os intervalos de confiança e HPD obtidos pelos MML e MB respectivamente abrangem o maior indivíduo dessas famílias, entretanto só o

MB conseguiu obter uma estimativa pontual satisfatória para o indivíduo em questão (Tabela 4).

Em geral para a variável MMC a eficiência média obtida quando se utilizou a estimativa do MML é elevada, já que para as médias dos três e cinco melhores indivíduos ficou entre 91,5% - 102,9% e 89,1% - 96,2% respectivamente, valores esses muito próximos do valor esperado que é igual a 1. Por outro lado, quando usou-se o MB a eficiência média apresentou resultados satisfatórios para $\bar{\boldsymbol{y}}_{\mathbf{3}}$, mas não para $\bar{\boldsymbol{y}}_{\mathbf{5}}$ em que só um valor pode ser considerado admissível (88,0%) (Tabela 4).

As estimativas pontuais para Brix obtida pelo MML ficaram abaixo dos maiores indivíduos nas quatro amostras, mas destaca-se que nas amostras de 43 e 63 famílias estas estimativas estão muito perto dos valores analisados (Tabela 5). É importante destacar que todos os respectivos intervalos de confiança incluem os valores máximos das amostras. Coles (2001), De Haan e Ferreira (2006) argumentam que os resultados obtidos pela metodologia da aproximação normal são poucos satisfatórios ao serem comparados com outras metodologias, entretanto, para as variáveis analisadas apresentam um comportamento similar à metodologia do Perfil de Verossimilhança.

As estimativas pontuais e os intervalos HPD obtidos por meio do MB são plausíveis, uma vez que todos os intervalos abrangem o valor extremo das amostras estudas, e as estimativas pontuais das amostras 30, 43 e 63 ultrapassam o maior indivíduo em cada uma destas (Tabela 5).

Para a variável Brix, a eficiência média obtida é relativamente alta, já que independentemente da estimativa (MML ou MB) e a média ($\bar{y}_3$ ou $\bar{y}_5$) utilizadas para seu cálculo, os resultados estão muito próximos ao valor esperado de 100% (98,5% - 101,5%). Deve-se destacar que para a amostra n = 63 a eficiência média obtida por meio da estimativa do MML e com $\bar{y}_3$ e $\bar{y}_5$, é igual a 100%, corroborando a aplicabilidade da teoria de valores extremos no melhoramento genético de plantas.

Com os dados reais, segundo a metodologia Q. MCMC, os comportamentos (Tabelas 4 e 5) da eficiência seletiva em função do aumento do número de famílias avaliadas são apresentados nas Figuras 2 e 3. A projeção da curva para além das 63 famílias experimentais foi realizada via interpolação harmônica usando o decréscimo na taxa de incremento na variável resposta (eficiência) associado à taxa de acréscimo na variável regressora (tamanho amostral ou número de famílias).

Tabela 5 - Comparação das estimativas pontuais e intervalos de confiança dos níveis de retorno com os valores máximos das amostras, e resultados das acurácias das médias dos três e cinco melhores indivíduos amostrais da variável Brix

Nov. Fam.	Máx	$\overline{y_3}$	$\overline{y_5}$	Métodos	Estimativa do máximo para o caráter	Intervalos	Eficiência Média	
							$\overline{y_3}$	$\overline{y_5}$
n = 20	18,97	18,85	18,72	Ve. P	18,562	(18,316; 20,014)	101,5%	100,9%
				A.normal	18,562	(18,053; 19,071)	101,5%	100,9%
				Q.MCMC	18,859	(18,238; 20,041)	99,9%	99,3%
n = 30	18,97	18,93	18,83	Ve. P	18,695	(18,684; 20,450)	101,3%	100,7%
				A. normal	18,695	(18,181; 19,210)	101,3%	100,7%
				Q.MCMC	19,011	(18,404; 20,220)	99,6%	99,1%
n = 43	18,97	18,94	18,91	Ve. P	18,789	(18,779; 20,818)	100,8%	100,6%
				A. normal	18,789	(18,247; 19,33)	100,8%	100,6%
				Q.MCMC	19,117	(18,523; 20,361)	99,1%	98,9%
n = 63	18,97	18,95	18,92	Ve. P	18,868	(18,826; 21,184)	100,4%	100,3%
				A. normal	18,868	(18,285; 19,450)	100,4%	100,3%
				Q.MCMC	19,208	(18,627; 20,484)	98,6%	98,5%

Máx.: Valor máximo - $\bar{y}_3$:Média dos três melhores indivíduos - $\bar{y}_5$:Média dos cinco melhores indivíduos - Ve. P: Perfil de Verossimilhança - A. Normal: Aproximação Normal - Q.MCMC: Quantis de Monte Carlo via Cadeias de Markov

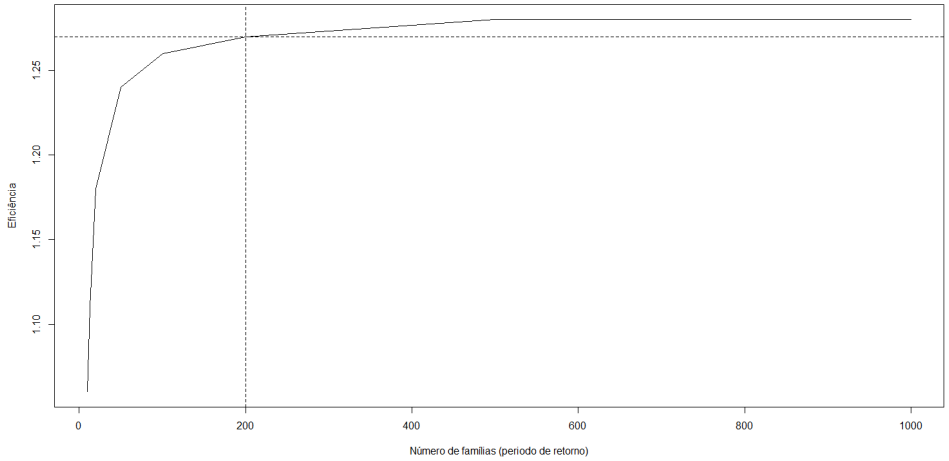

Figura 2 - Comportamentos da eficiência seletiva em função do aumento do número de famílias avaliadas para MMC.

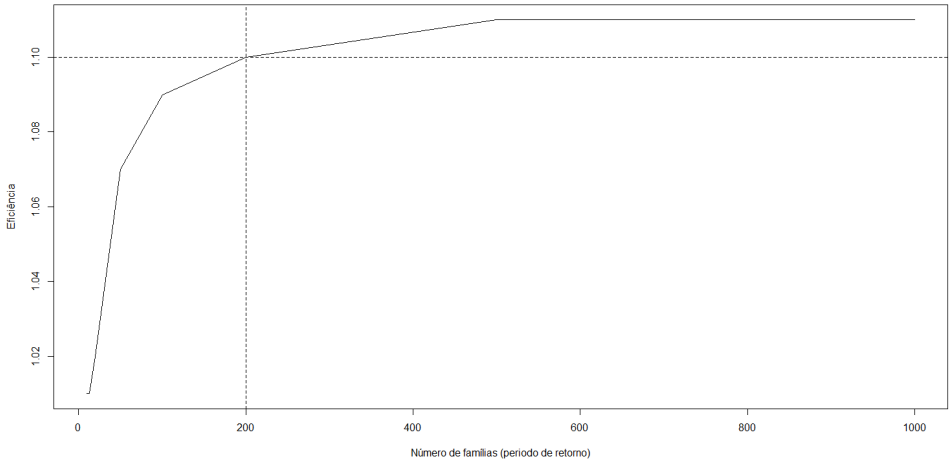

Figura 3 - Comportamentos da eficiência seletiva em função do aumento do número de famílias avaliadas para Brix.

Verifica-se que a eficiência (em relação ao uso de apenas 20 famílias) atingiu valores assintóticos em torno de 1,29 para a MMC e em torno de 1,10 para a Brix, associados a 200 famílias. Constata-se, portanto, que com o aumento do número de famílias há maior possibilidade de ganho genético para os caracteres em estudo. Castro *et al.* (2015) relatam que com um número grande de famílias avaliadas, há maior probabilidade de encontrar

clones promissores. Dessa forma, assim como na simulação estocástica, pode-se recomendar a avaliação de 200 famílias em cada ciclo seletivo.

3.3 Seleção de famílias pela capacidade de geração de indivíduos superiores

As populações simuladas podem ser tomadas como se fossem diferentes famílias e os resultados da Tabela 2 podem ser usados para aferir uma metodologia para classificar as famílias ou progênies pela capacidade de geração de indivíduos superiores ou excepcionais e informar os tamanhos amostrais a serem praticados em cada família para capturar esses indivíduos. Assim, conforme a Tabela 6, as famílias de códigos 1000 e 500 seriam selecionadas pela maior capacidade de geração de indivíduos extremos. E para recuperar um indivíduo extremo com valor 27, a família 100 demandaria tamanho amostral 200, a família 200 demandaria tamanho amostral 100, a família 500 demandaria tamanho amostral em torno de 50 e a família 1000 demandaria tamanho amostral 20 (Tabela 6).

Tabela 6 - Seleção pela maior capacidade de geração de indivíduos extremos para o caráter produtividade em toneladas de açúcar por hectare (TPH)

Código Família	n_{ind}	Valor máximo
20	20	20.03
20	100	22.43
20	200	23.37
50	20	22.17
50	100	24.62
50	200	25.38
100	20	23.53
100	100	25.99
100	200	26.6
200	20	24.73
200	100	27.18
200	200	27.61
500	20	26.08
500	100	28.49
500	200	28.69
1000	20	26.96
1000	100	29.32
1000	200	29.34

A seleção de famílias é fundamental no melhoramento vegetal, pois possibilita capitalizar as variações entre famílias que são de origem genética (CASTRO *et al.*, 2015; SILVA *et al.* 2015; ZHOU e MOKWELE, 2015). Esta é uma estratégia amplamente utilizada em associação com a seleção individual direcionada para clonagem nos programas de melhoramento de espécies de propagação assexuada, como o eucalipto e a cana-de-açúcar. De acordo com Castro *et al.* (2015), a seleção de famílias em cana-de-açúcar será importante, principalmente para caracteres quantitativos em nível de planta, uma vez que a

herdabilidade de família tende a ser maior do que a herdabilidade individual. Ademais, pela seleção de famílias de alto valor genético aumenta-se a chance de encontrar indivíduos excepcionais em suas progênies (RESENDE e BARBOSA, 2005; CASTRO *et al.*, 2015). Assim, os procedimentos apresentados aqui pela primeira vez na literatura permitem a otimização da seleção de famílias pela capacidade de geração de indivíduos superiores em um programa de melhoramento de plantas.

A metodologia mostrou-se funcional e é fundamentada da seguinte forma. Uma base de dados experimentais referentes à avaliação de famílias, mediante o uso de uma distribuição de valor extremo para predição do máximo das distribuições dos indivíduos, permite a previsão do comportamento da eficiência seletiva para os máximos associados a vários tamanhos de famílias e de populações experimentais. Isso possibilita ao melhorista a otimização da experimentação no melhoramento visando à seleção de indivíduos extremos. Para essas previsões, emprega-se o *período de retorno* associado à ocorrência de um nível do evento raro típico da distribuição ajustada. No caso, o período de retorno é interpretado como o tamanho amostral necessário para a ocorrência de determinado nível de retorno do evento raro (valor extremo com sua magnitude).

Considerando esses aspectos, uma boa opção prática seria a avaliação de 200 famílias com 100 indivíduos, perfazendo um total de 20000 indivíduos. Os modelos Gumbel e Weibull mostraram-se adequados para analisar a massa média de colmos (MMC) e teor de Brix (B%), independentemente do número de famílias na amostra ($n \geq 20$ famílias), sendo que a Gumbel se mostrou adequada apenas nos casos de números de famílias muito pequenos. Assim, recomenda-se a Weibull para inferências práticas. A funcionalidade genérica da distribuição Weibulll para a predição de valores extremos tem sido confirmada recentemente (GARDES e GIRARD, 2016).

Conclusões

De modo geral, a metodologia de valores extremos é adequada para classificar as famílias ou progênies pela capacidade de geração de indivíduos superiores ou excepcionais e informar os tamanhos amostrais a serem praticados em cada família para capturar esses indivíduos. Assim, é um processo de melhoramento genético do quantil máximo em programas de seleção em plantas.

Agradecimentos

Aos dois revisores e editores pelos comentários e sugestões.

ESCOBAR, J. A. D.; RESENDE, M. D. V.; AZEVEDO, C. F.; SILVA, F. F.; BARBOSA, M. H. P.; NUNES, A. C. P.; ALVES, R. S.; NASCIMENTO, M. Extreme value theory and sample size for the genetic improvement of the maximum quantile in plants. *Rev. Bras. Biom*., Lavras, v.36, n.1, p.108-127, 2018.

- ABSTRACT: *This study aimed to propose and evaluate a statistical methodology to improve the extreme value of the distributions. Such an approach is based on the upper quantiles of GEV (Generalized Extremes Values Distribution) of individual genotypic values between and within families. From real and simulated data sets from sugarcane families, generalized extreme value distributions (Gumbel, Fréchet and Weibull) were fitted to the maximum of each family. Stochastic*

simulations and experimental data resampling consistently indicated that the evaluation of 200 families is enough to maximize the efficiency in order to select extreme individuals. Weibull distribution fitted best and indicated an increase in selection efficiency is about 1.10 (gain of 10%) when going from 20 to 100 individuals per family and 1.12 (gain of 2%) when going from 100 to 200 individuals. These numbers are approximately constant regardless of the number of evaluated families. A good practical option would be the evaluation of 200 families with 100 individuals, in a total of 20,000 individuals. The methodology is also suitable to classify the families or progenies ability to generate exceptional individuals and inform the sample sizes to be practiced in every family to capture these individuals.

- KEYWORDS: *Probability distributions; vegetative propagation; outstanding individual; family size; number of families.*

Referências

BESAG, J. Markov Chain Monte Carlo for Statistical inference. *Center for Statistics and the Social Sciences.* Working Paper, n.9, 2001.

BOVIER, A. *Extreme values of random processes -Lecture Notes* -. Bonn: Institut für Angewandte Mathematik, 2010. 97p.

CAI, Y. Extreme value prediction via a quantile function model. *Coastal Engineering*, v.77, p.91-98, 2013.

CASTILLO, E.; HADI, A. S.; BALAKRISHNAN, N.; SARABIA, J. M. *Extreme Value and Related Models with Applications in Engineering and Science*. New Jersey: Wiley, 2005. 353p.

CASTRO, R. D.; PETERNELLI, L. A.; RESENDE, M. D. V.; MARINHO, C. D.; COSTA, P. M. A; BARBOSA, M. H. P; MOREIRA, E. F. A. Selection between and within full-sib sugarcane families using the modified BLUPIS method (BLUPISM). *Genetics and Molecular Research*, n.15, v.1, 2015.

CHERNOZHUKOV, V; FERNÁNDEZ-VAL. I. Inference for extremal conditional quantile models, with an application to market and birthweight risks. *The Review of Economic Studies*, v.78, n.2, p.559-589, 2011.

COLES, S. *An introduction to statistical modeling of extreme values.* Heidelberg: Springer, 2001. 205p.

DE HAAN, L.; FERREIRA, A. *Extreme value theory: An introduction.* New York: Springer, 2006. 417p.

EFRON, B.; TIBSHIRANI, R. J. *An introduction to the bootstrap*. New York: Chapman e Hall, 1993. 436p.

GARDES, L. GIRARD, S. On the estimation of the functional Weibull tail-coefficient. *Journal of Multivariate Analysis*, v. 146, p.29-45, 2016.

GILLELAND, E.; KATZ, R. W. New software to analyze how extremes change over time. *Eos*, n.92, v.2, p.13-14, 2011.

GUMBEL, E. J. *Statistics of extremes*. New York: Dover Publ., 2004. 371p.

JENKINSON, A. F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, v.81, p.159-171, 1955.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. Available at: https://www.R-project.org/.

RESENDE, M. D. V. Delineamento de experimentos de seleção para maximização da acurácia seletiva e do progresso genético. *Revista Arvore*, v.19, n.4, p.479-500, 1995.

RESENDE, M. D. V. Inferência Bayesiana e simulação estocástica (amostragem de Gibbs) na estimação de componentes de variância e de valores genéticos em plantas perenes. *Embrapa Florestas. Documentos, 46*. Colombo: Embrapa Florestas, 2000. 68 p.

RESENDE, M. D. V.; BARBOSA, M. *Melhoramento genético de plantas de propagação assexuada*, Colombo: Embrapa Florestas, 2005. 121p.

SCHAUMBURG, J. Predicting extreme value at risk: Nonparametric quantile regression with refinements from extreme value theory. *Computational Statistics & Data Analysis*, v.56, n.12, p.4081-4096, 2012.

SILVA, F. L.; BARBOSA, M. H. P.; RESENDE, M. D. V.; PETERNELLI, L. A.; PEDROSO, C. A. Efficiency of selection within sugarcane families via simulated individual BLUP. *Crop Breeding and Applied Biotechnology*, v.15, p.1-9, 2015.

SMITH, R. L. Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, v.72, p.67-90, 1985.

SPROTT, D. A. *Statistical inference in science*. New York: Springer, 2000. 229p.

STEPHENSON, A.; RIBATET, M. *evdbayes: Bayesian analysis in extreme value theory*. R package version 1.1-1, 2014 Available at: http://CRAN.R-project.org/package=evdbayes.

STEPHENSON, A. G. *evd: Extreme Value Distributions*. R News, n.2, v.2, p.31-32, 2002. Available at: http://CRAN.R-project.org/doc/Rnews/.

TRAUTMANN, H.; STEUER, D.; MERSMANN, O., BORNKAMP, B *truncnorm: Truncated normal distribution*. R package version 1.0-7, 2014. Available at: http://CRAN.R-project.org/package=truncnorm.

VON MISES, R. La distribution de la plus grande de n valeurs. [Reprinted (1954) in Selected Papers 11271-294]. *American Mathematical Society*, Providence, RI, 1936.

ZHOU, M.; MOKWELE, A. Family versus individual plant selection for stem borer (*Eldana saccharina*) resistance in early stages of sugarcane breeding in South Africa. *South African Journal of Plant and Soil*, 2015.

Recebido em 31.05.2016

Aprovado após revisão em 30.06.2017