UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS EXATAS, NATURAIS E DA SAÚDE
DEPARTAMENTO DE COMPUTAÇÃO

CAÍQUE DE OLIVEIRA DE SOUZA

FDQ: UM PACOTE R PARA GARANTIDA DA QUALIDADE DE EM BASES DE DADOS DE INVENTÁRIO FLORESTAL

ALEGRE – ES
2016

CAÍQUE DE OLIVEIRA DE SOUZA

FDQ: UM PACOTE R PARA GARANTIDA DA QUALIDADE DE EM BASES DE DADOS DE INVENTÁRIO FLORESTAL

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. D.Sc. Clayton Vieira Fraga Filho

ALEGRE - ES

2016

CAÍQUE DE OLIVEIRA DE SOUZA

FDQ: UM PACOTE R PARA GARANTIDA DA QUALIDADE DE EM BASES DE DADOS DE INVENTÁRIO FLORESTAL

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 24 de novembro de 2016.

COMISSÃO EXAMINADORA

Prof. D.Sc. Clayton Vieira Fraga Filho
Universidade Federal do Espírito Santo
Orientador

Prof. Dr. Fabrício Martins Mendonça
Universidade Federal do Espírito Santo

Prof. M.Sc Antônio de Almeida Barros Júnior
Universidade Federal do Espírito Santo

“Que vossos esforços desafiem as impossibilidades, lembrai-vos de que as grandes coisas do homem foram conquistadas do que parecia impossível. ”

Charles Chaplin

AGRADECIMENTOS

A priori gostaria de agradecer a Deus, que meu deu força ao longo de toda minha jornada durante a faculdade e nunca me deixou desistir de meus objetivos me mantendo firme quando mais precisei.

Agradeço minha família, em especial meu pais por me darem todo apoio durante o tempo que estudei, por nunca deixarem faltar nada e se esforçando para me manter durante meus estudos sempre estando ao meu lado para qualquer dificuldade que aparecesse e me dando força para que conseguisse chegar onde estou hoje. Agradeço também em especial ao meu tio que não está mais entre nós, mas sempre foi um exemplo de pessoa para mim, ele que sempre sonhou me ver formado espero que esteja lá do céu muito orgulhoso de mim.

Aos amigos que conheci durante o curso, em especial minha família em alegre Marília, Juliana (Shu), Maria de Lourdes (Lurdinha), Emannuel, vocês foram essenciais para mim durante todos esses anos. Agradeço também aos meus amigos Denis e Melissa por serem como irmãos em matérias, comilanças, festas, por sempre me escutarem e estarem ao meu lado nas horas boas e ruins. Contudo, não posso deixar de citar outros amigos que só agregaram felicidade na minha vida como Ana Paula, Leandro, Leonardo, Sílvia, Silas, Raphael, Carol, Tarine e também a todos do grupo “Social das Cachaças”. Quero que saibam que todos foram uma família para mim fora de casa durante todos esses anos, vocês foram fundamentais e só me trouxeram felicidades e tornaram minha vida muito melhor na universidade e foram um exemplo de companheirismo nas matérias realizadas juntos, nas festas, nas comidas, churrascos.

Ao meu orientador Prof. Clayton Vieira Fraga Filho pela orientação e condução deste trabalho, seus conhecimentos e contribuições para realização do mesmo foram fundamentais. Agradeço também por toda paciência e dedicação para comigo, sempre me atendendo nas madrugadas à dentro não se importando com a hora e me dando força para seguir em adiante e conseguir alcançar meus objetivos.

Por fim agradeço a todos os professores que me edificaram até onde estou, obrigado pelas aulas, pelos puxões de orelha, pela formação moral e teórica que me aprendi durante o curso e que me fizeram quem eu sou hoje.

SUMÁRIO

1 INTRODUÇÃO 14
1.1 Estrutura do Trabalho 19
2 OBJETIVOS 21
2.1 Objetivo geral 21
2.2 Objetivos específicos 21
3 REVISÃO DE LITERATURA 22
3.1 Qualidade de dados 22
3.1.1 Qualidade de dados no setor florestal 24
3.2 Métodos e frameworks para gestão da qualidade de dados 25
3.2.1 Framework de qualidade de dados (DQAF) 26
3.2.2 ISO/IEC 25012 27
3.2.3 ISO/IEC 25024 29
3.2.4 DMBOK 30
3.3 Inventário florestal 32
3.3.1 Inventário florestal contínuo 33
3.3.2 Tipos de origem de dados de inventário florestal 34
3.4 Estatística descritiva 35
3.5 Modelagem matemática 36
3.5.1 Modelo de regressão 37
3.5.2 Modelos de crescimento e produção 38
3.6 Problemas relacionados à Qualidade de Dados 45
3.7 Detecção e Correção de Erros 46
3.7.1 Dados longitudinais 47
3.8 Procedimentos de consistência 49
3.8.1 Procedimentos para árvore individual 49
3.8.2 Incremento diamétrico 53

3.9 Métricas de qualidade de dados de bases florestais 58
3.10 Repositório de Pacotes do R (CRAN) 59
4 Metodologia 60
4.1 Desenvolvimento do pacote para o R 61
4.2 Descrição dos dados 64
4.3 Processamento dos dados 65
5 RESULTADOS E DISCUSSÃO 66
5.1 Primeiro Estudo de Caso 66
5.1.1 Método de análise: check_measurement_ages 67
5.1.1 Método de análise: check_ages 69
5.1.2 Método de análise: check_parcel_different_spacing 69
5.1.3 Método de análise: check_clone_different_plot 69
5.1.4 Método de análise: check_dead_state 70
5.1.5 Método de análise: check_measurement_state 71
5.1.6 Método de análise: check_undefined_spacing 71
5.1.7 Método de análise: check_zero_measurement 72
5.1.8 Interseção de inconsistências 73
5.1.9 Método de análise: check_diametric_increase 75
5.2 Segundo Estudo de Caso 82
5.2.1 Método de análise: check_measurement_ages 82
5.2.2 Método de análise: check_ages 83
5.2.3 Método de análise: check_parcel_different_spacing 84
5.2.4 Método de análise: check_clone_different_plot 84
5.2.5 Método de análise: check_dead_state 84
5.2.6 Método de análise: check_measurement_state 84
5.2.7 Método de análise: check_undefined_spacing 85
5.2.8 Método de análise: check_zero_measurement 85

5.2.9 Interseção de inconsistências 85
5.2.10 Método de análise: check_diametric_increase 86
5.3 Comparativo dos Estudos de Caso 92
6 Conclusão 94
6.1 Trabalhos Futuros 94
7 REFERÊNCIAS BIBLIOGRÁFICAS 96
8 APÊNDICES 102
8.1 APÊNDICE A – Scripts De Execução Dos Métodos de Análises 102
8.1.1 Script realizado para obter a interseção de resultados 105
8.2 APÊNDICE B – Documentação do Pacote Forest Data Quality 106
8.3 APÊNDICE C – Documentação do Software Forest Data Quality Manager (FDQM) 126
8.3.1 Implementação do Sistema 126
8.3.2 Integração do FDQ com o sistema FDQM 137

LISTA DE FIGURAS

Figura 1 Mapa Mental da Estrutura do Trabalho 19
Figura 2. Estrutura em cascata de funcionamento do DQAF. 26
Figura 3. Área dividida em parcelas. 33
Figura 4. Apresentação de dados. 35
Figura 5. Tipos de Modelos de Crescimento e Produção. 39
Figura 6. Função de Densidade de Probabilidade Weibull, para diferentes valores de ɣ (forma) e β (escala). 41
Figura 7. Determinação do modelo de árvores individuais. 42
Figura 8. Origem dos dados de crescimento e produção. 43
Figura 9. Estados associados a covas, árvores e fustes em parcelas permanentes.44
Figura 10. Tipos de dados incompletos. 48
Figura 11 – Estados dos fustes representados por uma variável categórica. 49
Figura 12 Processo de Incremento Diamétrico. 54
Figura 13 Distribuição de classe por produtividade igual 32 em 24 meses. 55
Figura 14 Distribuição de classe por produtividade igual 26 em 24 meses. 55
Figura 15 Distribuição de classe por produtividade igual 20 em 24 meses. 55
Figura 16 Gráfico de mudança de classe da área com produtividade igual a 32 em 36 meses. 56
Figura 17 Gráfico de mudança de classe da área com produtividade igual a 26 em 36 meses. 56
Figura 18 Gráfico de mudança de classe da área com produtividade igual a 20 em 36 meses. 57
Figura 19 Fluxo de execução do trabalho. 60
Figura 20 Implementação das funções no R. 62
Figura 21 Diagrama de atividades para submissão do FDQ ao CRAN. 64
Figura 22 Exemplo de base não pareada. 66
Figura 23 Exemplo de base de dados pareada. 67
Figura 24 Resultados da avaliação do DAP. 68
Figura 25 Resultados da avaliação HT. 68
Figura 26 Resultado analise check_ages. 69
Figura 27 Resultados da análise check_dead_state. 70
Figura 28 Resultados da análise check_measurement_state. 71

Figura 29 Resultados análise check_undefined_spacing. 72
Figura 30 Resultado análise check_zero_measurement. 72
Figura 31 Incremento Diamétrico da parcela 356 aos 12 meses. 77
Figura 32 Incremento Diamétrico da parcela 356 aos 24 meses. 78
Figura 33 Incremento Diamétrico da parcela 356 aos 36 meses. 78
Figura 34 Incremento Diamétrico da parcela 356 aos 48 meses. 79
Figura 35 Incremento Diamétrico da parcela 356 aos 60 meses. 80
Figura 36 Saltos de classes parcela 356. 81
Figura 37 Resultado da análise check_measurement_ages para variável DAP. 82
Figura 38 Resultado da análise check_measurement_ages para variável HT. 83
Figura 39 Resultado da análise check_ages. 83
Figura 40 Resultado da análise check_measurement_state. 84
Figura 41 Incremento Diamétrico para parcela 149 aos 36 meses. 87
Figura 42 Incremento Diamétrico para parcela 149 aos 24 meses. 87
Figura 43 Incremento Diamétrico para parcela 149 aos 48 meses. 88
Figura 44 Incremento Diamétrico para parcela 149 aos 60 meses. 88
Figura 45 Incremento Diamétrico para parcela 149 aos 72 meses. 89
Figura 46 Incremento Diamétrico para parcela 149 aos 84 meses. 89
Figura 47 Incremento Diamétrico para parcela 149 aos 96 meses. 90
Figura 48 Saltos de classes da parcela 149. 91
Figura 49 Exemplo de base resultado após interseção 105

LISTA DE TABELAS

Tabela 1. Principais modelos em nível de povoamento 40

LISTA DE QUADROS

Quadro 1 Exemplo de árvore com DAP e HT=0 e estado diferentes de M, F, A.....50
Quadro 2 Exemplo de árvores com DAP2 < DAP1 entre medições .. consecutivas i e i+1, sendo i o ano da medição. 50
Quadro 3 Exemplo de número de fustes mortos (estado = M) ao longo das medições.....51
Quadro 4 Exemplo de número de clones em parcela. 51
Quadro 5 Exemplo de clone com parcela indefinida. 51
Quadro 6 Exemplo de parcela com espaçamento indefinido. 52
Quadro 7 Exemplo de parcela com espaçamento diferente.....52
Quadro 8 Exemplo de árvores que não atendam IDADE2 – IDADE1=1.....53
Quadro 9 Descrição da função de verificação de DAP. 63
Quadro 10 Interseção dos resultados da função check_measurement_ages.....73
Quadro 11 Interseção dos resultados da função check_ages.....74
Quadro 12 Interseção dos resultados da função check_ages.....74
Quadro 13 Interseção dos resultados da função check_measurement_state.....74
Quadro 14 Interseção dos resultados da função check_undefined_spacing 75
Quadro 15 Interseção dos resultados da função check_measurement_state.....75
Quadro 16 Interseção dos resultados da função check_measurement_ages.....86
Quadro 17 Interseção dos resultados da função check_ages.....86
Quadro 18 Interseção dos resultados da função check_measurement_state.....86
Quadro 19 Comparativo dos estudos de caso realizados 92

RESUMO

Com a importância que a informação possui na sociedade atual, empresas e instituições têm feito uso da mesma em seu desenvolvimento, tomada de decisão entre outros propósitos. Eis então que nasce uma preocupação com a qualidade dos dados que compõe essas informações, ou seja, surge a dúvida sobre o quão confiáveis, consistentes, válidos são esses dados. Na área florestal, os dados compõem desde inventários florestais até modelos de crescimento e produção, cujo os mesmos permitem inferir sobre condições futuras, sendo um dos componentes de um sistema de modelagem de povoamentos florestais, servindo para fins como tomadas de decisão, corte, plantio, aferição de resultados econômicos. Tais atividades são de extrema importância na área florestal, portanto é inegável a necessidade de que esses dados possuam uma boa qualidade. Neste trabalho, foi implementado um pacote de funções para a linguagem R, que recebeu o nome de *Forest Data Quality* (fqd), o mesmo capaz de realizar diversas analises em bases de dados florestais, seja ela provinda de um inventário florestal ou de um modelo de crescimento, o intuito do pacote é encontrar possíveis erros relacionados a qualidade desses dados com foco na consistência, completude e pontualidade dos dados. Para comprovar a eficácia do FDQ foram realizados dois estudos de caso em diferentes bases de dados, constatou-se que os resultados encontrados pelo *Forest Data Quality* foram fundamentais para demonstrar anomalias que afetavam diretamente a qualidade dos dados da base.

Palavras Chave: Qualidade de Dados, Inventário Florestal, Modelos de Crescimento e Produção.

1 INTRODUÇÃO

Atualmente os dados tem se tornado um dos mais importantes ativos de empresas, instituições governamentais e de ensino, sendo esta, parte de inúmeras bases de dados em diversas áreas como: engenharia, biologia, ciências florestais, computação, finanças, entre outras. Campos (2005) afirma que muitas vezes, a qualidade de dados da organização pode ser considerada uma vantagem competitiva, ou mesmo o núcleo do processo de desenvolvimento organizacional.

Tais dados desempenham um papel fundamental na sociedade da tecnologia e comunicação (TIC), pois eles estão presentes em aplicativos sociais, governamentais e empresariais, sendo assim, fundamentais em todas as relações entre governos, empresas e cidadãos. Batini e Scannapieca (2006) afirmam que a qualidade dos dados nos processos decisórios e operacionais é reconhecida por várias instituições e organizações. Além disso, Araújo e Batista (2007) relatam que a probabilidade de tirar vantagem dos recursos de informação é maior quando existe um conhecimento prévio do quão confiáveis são os dados utilizados como entrada. Informação de má qualidade provoca ineficiência operacional, compromete a tomada de decisão por levar a conclusões equivocadas, abala a confiança dos clientes, além de desgastar seu relacionamento com a empresa (LOSHIN, 2001).

O conceito de qualidade de dados indica o quanto as fontes de dados atendem às demandas de informação das áreas de negócio. Isto envolve conhecer a qualidade de dados atual, determinar a qualidade de dados desejada e ajustar a qualidade atual em função da desejada (TEIXEIRA FILHO, 1999). Qualidade de dados remete a saber o quão consistentes, completos e pontuais os dados estão (SEBASTIAN-COLEMAN, 2013).

Sebastian-Coleman (2013) define consistência como o grau em que os dados estão em conformidade com um conjunto equivalente de dados, normalmente produzido sob condições semelhantes, ou um conjunto produzido pelo mesmo processo ao longo do tempo. O autor ainda complementa dizendo que a característica de completude implica ter todas as partes necessárias ou adequadas dos dados. Em muitas empresas, as expectativas sobre os dados não podem ser atendidas porque são baseadas no pressuposto de que certos dados já existem, quando na verdade não existem, ou que dados existem numa forma utilizável, mas isso nem sempre é a realidade, uma vez que partes do conjunto, essenciais para a geração de informação

para a tomada de decisão não estão disponíveis. Ainda que se possa contar com dados consistentes e completos, espera-se que os mesmos estejam disponíveis para utilização no tempo certo, ou seja, no momento em que sejam demandados por seus usuários.

Por fim, Sebastian-Coleman (2013) apresenta uma característica fundamental, a pontualidade, que tem sido definida como o grau em que os dados representam a realidade a partir do ponto desejado no tempo. No que diz respeito ao processamento, a pontualidade está associada com à disponibilidade de dados, o grau em que os clientes têm os dados necessários no momento certo. Segundo Fraga Filho (2016) caso ocorra a indisponibilidade, os clientes que demandam a informação podem ficar frustrados causando estresse no ambiente organizacional e até mesmo levando membros da equipe a criarem dados irreais, ou usar métodos não recomendados para sua geração.

A ausência de fatores como consistência, completude e pontualidade acarretam em uma série de problemas que podem levar a sérios prejuízos. Inmon (1997) afirma que dados de má qualidade podem ser os motivos de falhas nos processos de negócio e custos relacionados às áreas organizacionais, podendo até haver perda parcial ou total de clientes.

Lima (2010) destaca a importância dada à qualidade de dados, quando relata que o Fundo Monetário Internacional (FMI) disponibiliza um modelo para aprimoramento e avaliação das: *Data Quality Assessment Framework* (DQAF), que fornece uma estrutura para avaliar a qualidade dos dados, comparando práticas estatísticas dos países com as melhores práticas, incluindo as metodologias internacionalmente aceitas. O *Data Warehousing Institute* em um relatório de 2002 sobre a qualidade de dados mostra que há uma diferença significativa entre a percepção e a realidade sobre a qualidade dos dados em muitas organizações, e que os problemas relacionados à falta de qualidade de dados custaram a empresas da América do Norte 600 milhões de dólares em um ano.

Araújo e Batista (2007) afirmam que sem informações de qualidade, não se consegue atender clientes de forma correta, tomar decisões adequadas ou obter vantagens de novas tecnologias. Segundo Kahn e Strong (1998) informação de qualidade é aquela que atende às suas especificações ou requisitos, portanto, a qualidade da informação é a característica de a informação atender ou exceder às

expectativas dos usuários. Huang et al (1999) definem a qualidade da informação como sendo a informação que é adequada para o uso por seus usuários.

Uma vez salientada a importância da qualidade de dados na tomada de decisão podemos destacar uma das técnicas utilizadas para realizar a mesma, uma delas é a modelagem matemática, que de acordo com Aguirre (2007) é a área do conhecimento que estuda maneiras de desenvolver e implementar modelos matemáticos de sistemas reais. Ou seja, tenta-se mapear a realidade através de uma linguagem matemática, geralmente através de equações. Aguirre (2007) destaca a utilização de modelos matemáticos ao longo da história para diversos fins, como, por exemplo: entender e explicar fenômenos observados na natureza quanto em sistemas sociais, biomédicos, predição, simulação e treinamento.

Tendo em vista a área florestal, um dos elementos essenciais do manejo florestal é a predição ou projeção aplicada em modelos de crescimento ou da produção, tais modelos que de acordo com Campos e Leite (2013) permitem inferir sobre condições futuras, sendo um dos componentes de um sistema de modelagem de povoamentos florestais. Tais modelos podem ter resultados obtidos de forma determinística ou estocástica. Aguirre (2007) salienta que modelos determinísticos são aqueles que variáveis e parâmetros são tratados de forma determinística e não como variáveis aleatórias. Portanto, os resultados serão obtidos exatamente a partir das condições ou valores iniciais. Por outro lado, quanto a modelos estocásticos, afirma que são aqueles que lidam com diversas fontes de incerteza presentes em qualquer situação real por meio do uso de variáveis aleatórias (AGUIRRE, 2007).

Modelos de predição são amplamente utilizados na agricultura, que é uma importante atividade para o crescimento econômico brasileiro, a mesma contribuiu com 21,46% do produto interno bruto do Brasil em 2015 (CEPEA 2015). As atividades agrícolas abrangem diferentes tipos e conjuntos de dados de uma variedade de fontes heterogêneas, em particular os mais comuns são os dados regulares e geoespaciais (MALAVERRI; MEDEIROS 2012). Com isso há a necessidade de conhecer, se o nível de qualidade dos dados utilizados é adequado, pois os mesmos serão usados em técnicas como análise de cultura. Além disso outras utilizações de tais dados são em aplicações e técnicas estatísticas para produção agrícola, tais como colheita de produção, taxa de semeadura, percentagem de plantio.

Nas ciências florestais, os dados compõem inventários florestais, que são fontes de informações sobre as características quantitativas e qualitativas da floresta e de muitas outras características das áreas as quais a floresta está se desenvolvendo. Os dados que compõem inventário florestal são provenientes de coleta, a mesma é fundamental, pois todo o restante do inventário se baseia nos dados coletados. Os dados devem ser precisos nos inventários, pois é através deles que se garante o retorno sobre o investimento. Dados errados podem conduzir a formulação de um planejamento equivocado (ISLAM et al., 2009). Duvemo (2009) afirma que a precisão de dados de inventário florestal afeta na qualidade de gestão de efeitos sobre empreendimentos florestais, pois leva a tomadas de decisões erradas quanto as estratégias de manejo e planejamento incorretos.

De acordo com Soares et al. (2006) um inventário florestal completo pode fornecer diversas informações entre elas: estimativas de área, descrição topográfica, mapeamento de propriedade, descrição de acesso, facilidade de transporte de madeira, estimativa da quantidade e qualidade de diferentes recursos florestais e estimativa de crescimento. Contudo, o inventário é a fonte de modelos de crescimento e produção florestal. Fraga Filho (2016) comenta que tais modelos permitem simular a dinâmica natural do povoamento, fornecendo resultados fundamentais para uma série de tarefas dentro de empresas florestais, como: planejamento estratégico, planejamento de colheita, aferição de resultados econômicos entre outros.

Erros em bases de dados sempre existirão. Quanto a localização e a correção dos mesmos, Batini e Scannapieca (2006) dizem que este processo é útil toda vez que os dados foram coletados a partir de fontes passíveis de erros, por exemplo, aqueles em que a entrada manual foi realizada, ou adquiridos a partir de fontes cuja confiabilidade não é totalmente garantida e sofre pressão para cumprimento de prazos, metas, dentre outros fatores.

Na área florestal, a localização e correção de erros antes que os dados componham modelos de crescimento ou de produção é fundamental, segundo Fraga Filho (2016) operações de consistência em bases de dados tem como objetivo encontrar desvios e problemas com a qualidade dos dados, buscando minimizar a inclusão de erros no processo de modelagem de crescimento e produção. Contudo, dentre os problemas relacionados a qualidade de dados na área florestal, destacam-se os relacionados à pontualidade, completude e consistência, a falta desses

requisitos pode acarretar diretamente nos processos operacionais e decisórios do manejo florestal, neste sentido, o processo de detecção e correção poderia contribuir para a área acadêmica e também para profissionais da área de ciências florestais, obtendo assim dados mais confiáveis e aptos para seus devidos fins.

1.1 Estrutura do Trabalho

Figura 1 Mapa Mental da Estrutura do Trabalho

O mapa mental apresentado na Figura 1 representa a estrutura geral do trabalho no qual cada ramo representa as seções principais do mesmo. Partindo da esquerda para direita, temos os objetivos onde são definidas as principais realizações do trabalho tais como a elaboração de uma ferramenta capaz de avaliar tanto bases de dados de inventário florestal quanto dados para entrada de modelos de crescimento e produção, além da realização de dois estudos de caso como forma de validação e verificação da eficácia da ferramenta.

O próximo passo é a revisão de literatura, onde são apresentados os conceitos nos quais fundamentou-se para realização do trabalho, quanto a qualidade de dados podemos ver a estrutura hierárquica dos frameworks e métodos existentes atualmente como DQAF, DMBOK, ISO/IEC 25012 e 25024, posteriormente temos a qualidade no setor florestal, o que dados de má qualidade podem acarretar na área e seus contras na tomada de decisão e nos processos decisórios e operacionais. Um ramo que representa bases de dados apresenta os tipos de dados que iremos trabalhar, os mesmos partem de dados de inventário florestal para modelos de crescimento e produção que se subdividem em MAI, MDD, MNP que são tipos de modelos de crescimento e produção. Vale observar que na revisão de literatura também são apresentados os procedimentos de para detecção de inconsistências que farão parte da ferramenta, os mesmos foram propostos por Fraga Filho (2016) e Demolinari (2006).

Quanto a metodologia são apresentadas as maneiras de como a ferramenta será desenvolvida, o desenvolvimento do pacote na linguagem R e seus benefícios, a validação do mesmo através dos estudos de caso, posteriormente temos os resultados e discussão, que é a parte onde iremos apresentar o resultado do funcionamento da ferramenta aplicadas as bases de dados no estudo de caso, por fim chegamos à conclusão que iremos avaliar a idealização de nossa ferramenta, seus benefícios, sua utilidade para comunidade em geral de profissionais da área e o impacto da detecção de inconsistências no planejamento florestal. Nos apêndices são apresentados os scripts realizados para obtenção dos resultados apresentados neste trabalho e também a documentação da ferramenta e de um software futuro que servirá como interface gráfica da ferramenta e proporcionará um sistema de gestão de configuração das bases de dados trabalhadas.

2 OBJETIVOS

2.1 Objetivo geral

Desenvolver uma ferramenta para avaliar a qualidade de dados de bases de dados florestais.

2.2 Objetivos específicos

a. Avaliar a qualidade de dados de inventários florestais contínuos (IFC) quanto à consistência, completude com foco na pontualidade.

b. Avaliar a qualidade de dados utilizados como entrada nos modelos de crescimento e produção.

c. Aplicar a ferramenta desenvolvida na realização de dois estudos de caso envolvendo bases de inventário florestal pareadas e não pareadas.

3 REVISÃO DE LITERATURA

3.1 Qualidade de dados

Dados eletrônicos desempenham um papel fundamental na sociedade da tecnologia da informação e comunicação (TIC): eles são gerenciados por aplicativos empresariais e governamentais. Além disso estão presentes em todos os tipos de aplicativos na Web e são fundamentais em todas as relações entre governos, empresas e cidadãos. Uma vez que os dados se tornaram um ativo essencial da sociedade, a "qualidade" dos mesmos passou a ter efeitos críticos sobre cada negócio.

Batini e Scannapieca (2006) afirmam que a qualidade dos dados nos processos decisórios e operacionais é reconhecida por várias instituições e organizações, como por exemplo, na declaração da missão do *European Statistical System*: "Nós fornecemos à União Europeia e ao mundo com informações de alta qualidade sobre economia e a sociedade no nível Europeu, nacional e regional e tornamos a informação disponível a todos para fins de tomada de decisão, pesquisa e debate.". Afirmam ainda que a qualidade dos dados também é uma questão essencial para os processos operacionais das empresas e organizações. O *Data Warehousing Institute* em um relatório de 2002 sobre a qualidade dos dados mostra que há uma diferença significativa entre a percepção e a realidade sobre a qualidade dos dados em muitas organizações, e que os problemas de qualidade de dados custaram a empresas da América do Norte 600 milhões de dólares em um ano.

A qualidade dos dados é definida por dois fatores relacionados: o quão bem ele atende às expectativas dos consumidores de dados (o quão bem ele é capaz de servir aos propósitos de sua utilização ou utilizações) e quão bem ele representa os objetos, eventos e conceitos que é criado para representar. A fim de medir se os dados atendem às expectativas ou estão "apto para a utilização", expectativas e usos precisam ser definidos (SEBASTIAN-COLEMAN, 2013).

Na tecnologia da informação o desenvolvimento de metodologias para avaliação da qualidade dos bancos de dados se baseia nos princípios de gerenciamento da qualidade total.

O DQAF destaca em um conjunto de características de dados associados cinco dimensões da qualidade dos dados: integridade, oportunidade, validade, consistência

e integridade (MALAVERRI; MEDEIROS, 2012; SEBASTIAN-COLEMAN, 2013). Sebastian-Coleman (2013) apresenta cada definição:

> Completude: implica em ter todas as partes necessárias ou adequadas dos dados. A primeira condição é a existência de completude, afinal devem existir dados antes que possa ser classificado completo. Em muitas empresas, as expectativas sobre os dados podem não ser atendidas porque são baseadas no pressuposto de que certos dados já existem quando não existem, ou que os dados existem numa forma utilizável, mas isso não é a realidade. Um conjunto de dados é completo na medida em que ele contém variáveis (ou atributos) necessárias e um número suficiente de registros, e na medida em que são preenchidas de acordo com as expectativas dos consumidores de dados. Para um conjunto de dados para ser completo, pelo menos, três condições devem ser atendidas: o conjunto de dados deve ser definida de modo a incluir todos as variáveis desejados (width); o conjunto de dados deve conter a quantidade desejada de dados (depth); e as variáveis devem ser preenchidas até ao ponto desejado (density). Também pode ser entendido em relação ao processamento de dados, por exemplo, ter todos os dados necessários ou convenientes para iniciar o processamento e garantir que parte dos dados não foi retirado de um conjunto de dados quando ele é processado
>
> Pontualidade: tem sido definida como o grau em que os dados representam a realidade a partir do ponto desejado no tempo. No que diz respeito ao processamento, a pontualidade está associada com a disponibilidade de dados, o grau em que os clientes têm os dados necessários no momento certo.
>
> Validade: A validade é o grau em que os dados estão em conformidade com um conjunto de regras de negócio, às vezes expressa como um padrão ou representado dentro de um domínio de dados definido. Validade é diferenciada de precisão e exatidão, cada um dos quais requer comparação com objetos reais. Validade se limita a medir contra substitutos ou alguns valores alternativos de objetos reais. Estes substitutos podem ser instanciados como dados, permitindo validade para ser medida diretamente a partir de um conjunto de dados de uma forma que a precisão e exatidão não pode ser.
>
> Consistência: A consistência é o grau em que os dados estão em conformidade com um conjunto equivalente de dados, normalmente um conjunto produzido sob condições semelhantes, ou um conjunto produzido pelo mesmo processo ao longo do tempo. Medidas de consistência podem definidas em conformidade com as regras de negócio (dependências que

podem ser estabelecidas tecnicamente se a lógica pode ser claramente expressa), ou podem revelar padrões lógicos criados dentro dos dados que refletem as conexões entre as situações do mundo real que os dados representam.

Integridade: Representa o grau em que os dados estão em conformidade com regras de relacionamento de dados (como definido pelo modelo de dados) que se destinam a assegurar a apresentação completa, consistente e válida de dados que representam os mesmos conceitos, normalmente considerando outras dimensões.

3.1.1 Qualidade de dados no setor florestal

Dados de inventário florestal e aqueles necessários ao ajuste de modelos de crescimento e produção nos diferentes níveis devem apresentar tais características, ou a qualidade da informação produzida pode não representar qualquer aderência com a realidade das florestas plantadas.

A validade de dados deve ser considerada não apenas pontualmente em cada medição, mas de forma a manter a integridade entre medições ao longo do período de existência de uma parcela, e os indivíduos que a compõem. Sendo assim, do ponto de vista da completude, para modelos de crescimento e produção, completo é o dado caracterizado por medida repetidas do mesmo indivíduo.

A integridade deve ser mantida na medida em que os dados produzidos, por exemplo para as prognoses, apresentem a relação de origem-destino e sua rastreabilidade garantida. Os resultados de uma prognose devem ser rastreados para os modelos pelos quais foram produzidas, pelas estatísticas de ajuste, análises realizadas, até o nível da árvore individual para mais.

Uma característica fundamental é a pontualidade, sobretudo para o processamento, espera-se que o dado esteja disponível no momento em que seja necessário realizar um estudo de prognose para avaliação de disponibilidade de múltiplos produtos em uma idade futura. Caso ocorra a indisponibilidade, os clientes que demandam a informação podem ficar frustrados causando estresse no ambiente organizacional e até mesmo levando membros da equipe a criarem dados irreais, ou usar métodos não recomendados para sua geração.

Tanto a completude quanto a consistência devem ser avaliadas em cada medição e entre medições. A falta de informações sobre o DAP (diâmetro do tronco a 1,3m do solo) de uma ou mais árvores em algumas parcelas pode prejudicar o ajuste

de um modelo que seria promissor para determinado propósito. Contudo, mesmo na presença do valor número do DAP é necessário avaliar, ou consistir em sua validade frente ao comportamento esperado para esta variável. Faz-se ainda necessária a consistência destas variáveis do ponto de vista longitudinal, ou seja, entre o conjunto de medições realizadas.

Medições são realizadas anualmente por equipes nas parcelas que são demarcações de áreas em plantios comerciais de árvores, por exemplo, eucalipto. A ordem de entrada na parcela é definida pela ordem que uma árvore é mensurada e adicionada a sua respectiva parcela, caso tal ordem não seja respeitada pode acarretar em problemas como: uma árvore não ser colocada na parcela que deveria, gerando assim erros nas modelagens que utilizarão tais dados como crescimento alterado em determinada parcela. Marcatti (2013) argumenta que no inventário florestal, as parcelas devem ser visitadas em uma ordem predefinida, para serem então mensuradas, e que encontrar a rota que minimiza o custo de caminho traz benefícios óbvios ao rendimento do inventário, destacando-se a redução de custos.

Como é necessário coletar medidas das árvores individuais, as mesmas recebem um identificador utilizado e reconhecido pelo usuário na base de dados de inventário florestal. Contudo, pode ocorrer que ao realizar uma medida no ano 2, a equipe inicie pela fileira incorreta, identificando a última árvore como a primeira, o que pode gerar uma inconsistência no ajuste dos modelos, caso o problema não seja detectado antes. Por um erro cometido em diferentes medições, por exemplo, a inversão da ordem de entrada em uma parcela em diferentes períodos de inventário pode acarretar em comportamento decrescente do valor de HT das mesmas covas ao longo do tempo. Tal característica é fundamental para garantir a qualidade dos dados para ajuste de modelos de crescimento e produção.

3.2 Métodos e *frameworks* para gestão da qualidade de dados

Atualmente existem soluções para gerir qualidade de dados, que vão desde frameworks até mesmos manuais de como chegar em dados que tenham qualidade suficiente para atender seus propósitos. Dentre as soluções temos um modelo de aprimoramento e avaliação de informações disponibilizado pelo Fundo Monetário Internacional (FMI) chamado de *Data Quality Assessment Framework* (DQFA). Além disso existem normas ISO de certificação sobre qualidade de dados como a ISO IEC

25012 que tem como objetivo definir um modelo geral de qualidade de dados para os dados conservados em um formato estruturado dentro de um sistema de computador. Outra ISO relacionada a qualidade de dados é a IEC 25024, que atua como um complemento da ISO 25012, definindo medidas de qualidade de dados para obter quantitativamente a qualidade dos dados em termos de características definidas na ISO / IEC 25012 apresentada acima. Por fim uma solução alternativa é o DMBOK (*Data Management Body of Knowledge*), uma publicação que apresenta uma base de conhecimento para gestão de dados que tem como finalidade planejar, implementar e controlar atividades que apliquem técnicas de gerência de qualidade de dados para medir, avaliar, melhorar e garantir a adequação dos dados ao seu uso pretendido.

3.2.1 Framework de qualidade de dados (DQAF)

O DQAF é um framework de avaliação de dados que fornece uma estrutura para avaliar a qualidade dos dados, comparando práticas estatísticas dos países com as melhores práticas, incluindo as metodologias internacionalmente aceitas. O DQAF (2012), abrange cinco dimensões da qualidade (garantias de integridade, solidez metodológica, exatidão e confiabilidade, manutenção e acessibilidade de qualidade de dados) e um conjunto de pré-requisitos para a avaliação da qualidade dos dados como: se o ambiente é favorável as estatísticas, se os recursos são condizentes com as necessidades do software estatístico e se a qualidade é um dos pilares do trabalho estatístico. *DQAF Generic Framework* (2003) afirma que para cada dimensão, o DQAF identifica 3-5 elementos de boas práticas, e para cada elemento, vários indicadores relevantes. A Figura 2 apresenta os detalhes em uma estrutura de cascata sobre como o DQAF funciona.

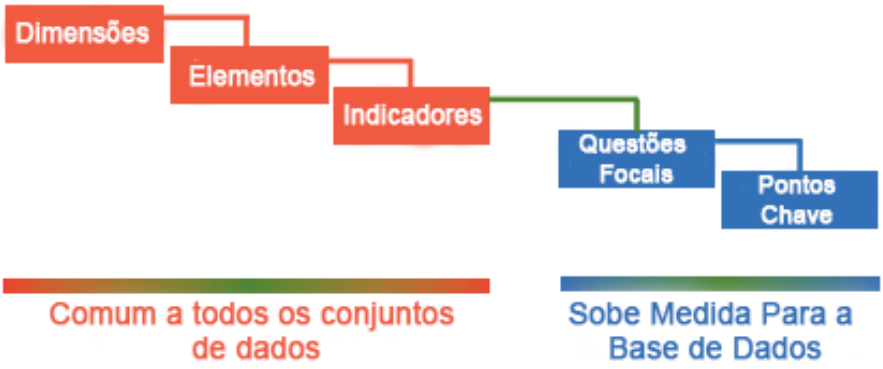


Figura 2. Estrutura em cascata de funcionamento do DQAF.

Fonte: Adaptado de DQAF Generic Framework (2003)

. A parte em vermelho da Figura 2 apresenta os pontos comuns a todos os conjuntos de dados, ou seja, pontos que serão avaliados para todo e qualquer tipo de dado que se deseja aplicar o framework. Primeiramente são avaliadas as dimensões em relação a este conjunto, após isso é realizada a identificação de boas práticas para cada elemento identificado pelas dimensões de qualidade, um exemplo de elemento que pode ser identificado é a avalição da fonte de dados.

Posteriormente vários indicadores são levados em questão, os mesmos representarão quantitativamente a qualidade de dados, um exemplo de indicador é a avalição dos dados de origem através de censos, pesquisas amostrais, e registros administrativos rotineiros avaliando, por exemplo, erros amostrais, erros de resposta.

A cobertura destas dimensões e desses pré-requisitos reconhece que a qualidade dos dados abrange características relacionadas à instituição ou sistema por trás da produção dos dados, bem como características do produto de dados individuais. O framework foi desenvolvido como uma metodologia de avaliação que visa proporcionar estrutura e uma linguagem comum para a avaliação da qualidade dos dados.

O DQAF facilita uma visão abrangente da qualidade dos dados, que reconheça inter-relações, incluindo compensações, entre elementos de qualidade e permite ênfases de variar entre as categorias de dados e usuários. Ele facilita o diálogo com as agências nacionais de estatística e as autoridades do país, bem como uma abordagem mais homogênea para avaliar a qualidade dos dados pela equipe do Fundo (DQAF, 2012).

3.2.2 ISO/IEC 25012

A ISO define um modelo geral de qualidade de dados para os dados conservados em um formato estruturado dentro de um sistema de computador, a mesma pode ser usada para estabelecer requisitos de qualidade de dados, definição de medidas de qualidade de dados, ou planejar e realizar avaliações de qualidade de dados (BENTANCOURTI; ROCHA, 2012). Alguns de seus objetivos são aqui apresentados:

a) Definir e avaliar os requisitos de qualidade de dados em processos de produção de dados, aquisição e integração.

b) Para identificar os critérios de garantia da qualidade dos dados, também úteis para reengenharia, avaliação e melhoria dos dados.
c) Para avaliar a conformidade dos dados com a legislação e / ou requisitos.

Bentancourti e Rocha (2012) afirmam que a norma ISO/IEC 25012 aborda especificamente a qualidade dos dados de forma multidimensional, segundo critérios classificados em duas categorias. A primeira categoria reúne critérios relativos à qualidade inerente dos dados, em que é possível atender uma demanda informacional a partir dos dados em si. A segunda, por outro lado, dependente do sistema, presume que a qualidade do dado é melhorada ou dependente do sistema computacional em que opera. No entanto, também há casos em que ambas as categorias estão presentes, ou seja, há demanda dos dados e existe uma dependência do sistema.

A ISO categoriza os atributos de qualidade em quinze características consideradas por dois pontos de vista: qualidade inerente aos dados e dependente do sistema. Os 15 critérios inerentes ao sistema, denominadas dimensões, são apresentados e definidos por Bentacourti e Rocha (2012):

> -Exatidão (*accuracy*): Grau no qual o dado representa o valor correto do atributo intencionado de um conceito ou evento em um contexto específico de uso.
>
> Completude (*completeness*): Grau ao qual o dado do sujeito associado com uma entidade tem valores para todos os atributos esperados e instâncias de entidades relacionadas em um contexto específico de uso.
>
> Consistência (*consistency*): Grau no qual o dado tem os atributos que são livres de contradição e são coerentes com outros dados em um contexto específico de uso.
>
> Credibilidade (*credibility*): Grau no qual o dado tem atributos que são considerados como verdadeiros e fidedignos pelos usuários em um contexto específico de uso.
>
> Atualidade (*currentness*): Grau no qual o dado tem os atributos que são do período correto em um contexto específico de uso.
>
> Os critérios inerentes e dependentes do sistema são:
>
> Acessibilidade (*accessibility*): Grau no qual o dado pode ser acessado em um contexto específico de uso, em particular por quem necessita do suporte de tecnologia ou de uma conFiguração especial decorrente de alguma necessidade especial ou incapacidade.
>
> Conformidade (*compliance*): Grau no qual o dado tem atributos que se aderem a normas, convenções ou regulamentações vigentes e regras

similares relacionadas com a qualidade de dados em um contexto específico de uso.

Confidencialidade (*confidentiality*): Grau no qual o dado tem os atributos que garantam que este é somente acessível e interpretável por usuários autorizados em um contexto específico de uso;

Eficiência (*efficiency*): Grau no qual o dado tem os atributos que podem ser processados e proporciona os níveis esperados de funcionamento (desempenho) utilizando as quantidades e os tipos de recursos apropriados em um contexto específico de uso.

Precisão (*precision*): Grau no qual o dado tem os atributos que são exatos ou que proporcionam a discriminação em um contexto específico de uso.

Rastreabilidade (*traceability*): Grau no qual o dado tem atributos que proporcionam um rasto de auditoria de acesso aos dados e de qualquer alteração ocorrida com os mesmos em um contexto específico de uso.

Entendibilidade (*undestandability*): Grau no qual o dado tem atributos que lhe permitem ser lidos e interpretados por usuários, e é expresso em linguagens apropriadas, símbolos e unidades em um contexto específico de uso.

Finalmente, os critérios dependentes do sistema são:

Disponibilidade (*availability*): Grau no qual o dado tem atributos que lhe permitem ser recuperados por usuários autorizados e/ou aplicações em um contexto específico de uso.

Portabilidade (*portability*) - grau no qual o dado tem os atributos que lhe permitem ser instalados, substituídos ou removidos de um sistema a outro conservando a qualidade existente em um contexto específico de uso.

Recuperabilidade (*recoverability*): Grau no qual o dado tem atributos que lhe permitem manter e conservar um nível especificado de operações e qualidade, mesmo em caso de falhas (do sistema), em um contexto específico de uso.

3.2.3 ISO/IEC 25024

A ISO tem como objetivo a medição da qualidade dos dados: Para isso fornece medidas, incluindo métodos de medição associados a elementos de mensuração de qualidade para as características no modelo de qualidade de dados. A mesma define medidas de qualidade de dados para medir quantitativamente em termos de

características definidas na ISO / IEC 25012. De acordo com a ISO 25024, as funções relacionadas a qualidade de dados presentes na mesma são:

- Um conjunto básico de medidas de qualidade de dados para cada característica.
- Um conjunto básico de entidades-alvo a que se aplicam as medidas de qualidade durante o ciclo de vida de dados.
- Uma explicação de como realizar a aplicação de medidas de qualidade de dados.
- Uma orientação para as organizações que definem as suas próprias medidas para os requisitos de qualidade de dados e avaliação.

Esta norma não define faixas de valores dessas medidas de qualidade aos níveis das taxas ou graus, porque esses valores são definidos para cada sistema, pela sua natureza, dependendo do contexto do sistema e as necessidades dos utilizadores. Ela pode ser aplicada a qualquer tipo de dados retidos num formato estruturado dentro de um sistema de computador utilizado para qualquer tipo de aplicação.

3.2.4 DMBOK

Ribeiro (2014) define o DMBOK como um documento da DAMA *International*, uma organização sem fins lucrativos e independente de fornecedores, constituída pela associação dos técnicos e profissionais dedicados a promover os conceitos e as melhores práticas em gestão, governança e qualidade de dados e informação, que apresenta um referencial que descreve e estrutura o processo de gestão de dados, indicando suas melhores práticas, princípios orientadores, definições padronizadas de resultados, papéis e terminologias, organizando a gestão de dados em funções de dados ou áreas de conhecimento.

Uma das melhores práticas presentes no DMBOK está relacionada ao gerenciamento da qualidade de dados. Barbieri (2013) salienta que o DMBOK descreve o objetivo da qualidade de dados como: planejar, implementar e controlar atividades que apliquem técnicas de gerência de qualidade de dados para medir, avaliar, melhorar e garantir a adequação dos dados ao seu uso pretendido. Além disso o Barbieri (2013) apresenta a estrutura da atividade de gestão de qualidade de dados e define cada segundo o DMBOK:

1. Desenvolver e promover aspectos de conscientização sobre qualidade de dados: entende-se que vender a importância da qualidade de

dados nas empresas. É preciso difundir a importância dos conceitos, seja por mecanismos diretos ou indiretos. Os diretos seriam através de palestras, consultorias ou assemelhados.

2. Definir requisitos de Qualidade de Dados: Os requisitos de qualidade de dados são definidos diretamente em função das necessidades da empresa. Barbieri (2013) apresenta os requisitos:

- Precisão (*accuracy*) ou como as "coisas"/entidades da vida real estão corretamente representadas.

- Completude (*completeness*): O quão completo estão os dados exigidos na execução daquele processo de negócio. (Todos os atributos? Faltam alguns? Todos os essenciais? Alguns acessórios?)

- Consistência (*consistency*): Se refere à integridade cruzada entre duas ou mais fontes que armazenam o mesmo dado. Há coerência entre esses dados que habitam fontes diferentes? A coerência existe no mesmo contexto ou em contextos diferentes?

- Atualidade (*currency*): O quanto os dados estão atualizados e representam o estado corrente e mais atual.

- Precisão numérica (*accuracy*): Representação de valores no grau de precisão necessária, como casas decimais para dados numéricos.

- Disponibilidade (*availability*): O dado é disponibilizado no momento de sua necessidade?

- Unicidade (*uniquiness*): O fato de haver representação única de certa entidade, sem ambiguidade ou sentidos diferentes.

3. Estabelecer processos de "*profiling*", análise e avaliação de Qualidade de Dados: refere-se à necessária fotografia inicial do estado dos dados de certa (s) área (s) de assunto ou domínio (s) da empresa. Deve ser uma das primeiras ações para se estabelecer as "*baselines*" dos processos de melhoria de qualidade de dados da empresa. Permite criar as primeiras métricas e a definição dos objetivos a serem alcançados em função delas.

4. Definir regras de negócios para Qualidade de Dados: Implica na análise das regras de negócios fundamentais dos processos e na descoberta dos dados que podem implicar em eventuais quebras de conformidade delas. Esses dados deverão ser observados na sua qualidade justamente para garantir a conformidade da regra com o processo.

5. Testar e validar os requisitos de Qualidade de Dados: Nesse ponto, o DMBOK sugere que haja processo de verificação inicial (data profiling, por exemplo) e verificação constante e recorrente, a fim de que os dados sejam sempre avaliados nos seus domínios de qualidade.

6. Definir e avaliar níveis de serviços de Qualidade de Dados: Nesse item, é sugerida a definição de níveis de serviços de qualidade de

dados, o que deverá ser garantido por medições e verificações constantes. Os níveis de SLA são o compromisso firmado sobre qualidade da área gestora dos dados com os seus usuários. Os itens subsequentes, relativos a medir e monitorar continuamente, gerenciar as pendências e corrigir os defeitos são consequências diretas desse compromisso de nível de serviços.

7. Medir e monitorar continuamente a Qualidade de Dados: Projetar e implementar procedimentos operacionais de Gerência de Qualidade de dados e monitorar os procedimentos operacionais e a performance da Gerência de Qualidade de Dados. São, no fundo, a materialização do processo de Garantia de Qualidade dos Dados. Todo processo deverá ser constantemente avaliado a fim de se obter melhorias nos seus procedimentos, políticas e resultados.

Por fim, através das práticas acima citadas se têm a materialização do processo de garantia de qualidade dos dados. Todo processo deverá ser constantemente avaliado a fim de se obter melhorias nos seus procedimentos, políticas e resultados.

3.3 Inventário florestal

O inventário florestal é o procedimento para obter informações sobre as características quantitativas e qualitativas da floresta e de muitas outras características das áreas sobre as quais a floresta está desenvolvendo. Segundo Kangas et al. (2006), o objetivo do inventário florestal é medir e estimar as características florestais de uma área pré-definida. Para Veiga (1984), os objetivos de um inventario florestal, podem variar. Eles serão definidos de acordo com o uso da área e os serviços ambientais. Entre eles: exploração florestal em grande, média e pequena escala, reserva florestal, bacias hidrográficas, manutenção da vida silvestre e recreação. Os inventários podem atender a interesses específicos de uma empresa florestal ou de uma instituição de pesquisa, visando uma determinada fazenda, uma parte de uma propriedade, ou um conjunto de propriedades.

Soares et al. (2006) afirma que um inventário florestal completo pode fornecer diversas informações entre elas:

a) Estimativa de área.
b) Descrição da topografia.
c) Mapeamento da propriedade.
d) Descrição de acessos.
e) Facilidade de transporte da madeira.

f) Estimativa da quantidade e qualidade de diferentes recursos florestais.

g) Estimativa de crescimento (se o inventário for realizado mais de uma vez).

O inventário florestal, é uma técnica importante, talvez até mesmo imprescindível ao bom conhecimento do potencial dos recursos existentes em uma determinada área. Com base nele, pode ser tomada decisões precisas e fidedignas para melhor conservação e manejo destes recursos (SANQUETTA et al., 2009).

3.3.1 Inventário florestal contínuo

Dentro os tipos de inventários florestais existentes, destaca-se o inventário florestal contínuo (IFC). Há algumas características que distinguem o inventário florestal convencional do IFC, segundo Campos e Leite (2013) enquanto o inventário florestal quase sempre usa parcelas temporárias, o IFC emprega parcelas permanentes. Conclui-se, portanto, que um inventário convencional tem como principal objetivo estimar a produção na época da medição (mensuração da árvore, coleta de valores como DAP que é o diâmetro da altura do peito, altura, idade, entre outros), enquanto um IFC se presta também para estimar mudanças ocorridas tanto na produção, como em área florestal, entre outras características. A Figura 3 apresenta a disposição de uma região com árvores separadas por parcela.

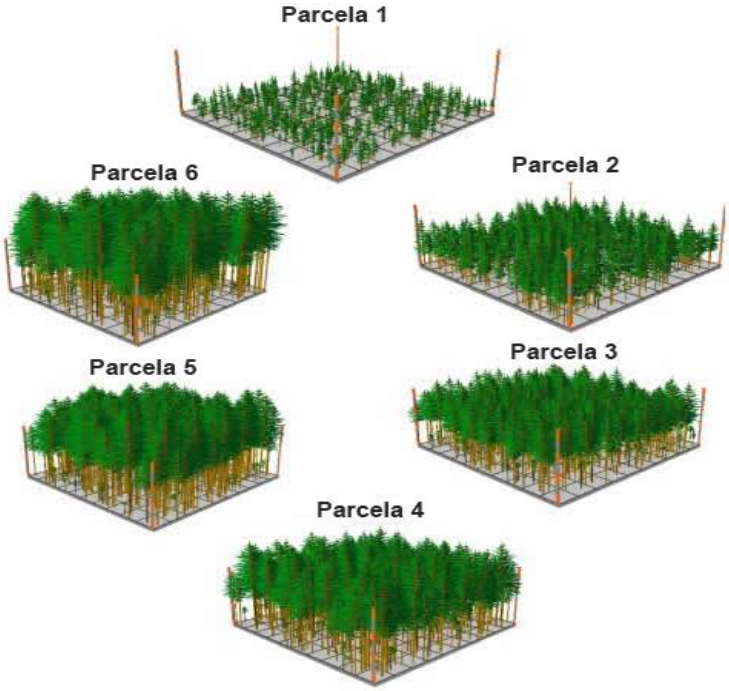


Figura 3. Área dividida em parcelas.

A Figura 3 representa conjuntos de parcelas compostas por diversas árvores, vale ressaltar que cada parcela tem uma única ordem de visita, essa ordem quando não é respeitada acarreta em muitos erros no modelo, relacionados a medição.

Com o uso do inventário florestal contínuo, podem-se detectar causas de mudanças ocorridas, pois, todas as árvores de parcelas permanentes são remedidas periodicamente. Sendo assim é possível quantificar mortalidade, sinistros florestais, ingresso e corte.

3.3.2 Tipos de origem de dados de inventário florestal

Dados de inventários florestais podem ser de diversos tipos e origens, Islam et al. (2009) afirma que os mesmos podem ser coletados por meio de amostragem de campo, mantendo os princípios estatísticos ou pela aplicação de procedimentos de amostragem subjetiva. Métodos de sensoriamento remoto, tais como o escaneamento a laser deverão se tornar padrão em um futuro próximo. Contudo, dados das árvores distribuídas pela rede de parcelas de um empreendimento florestal são do tipo alfanuméricos, ou seja, quantitativos como DAP (diâmetro a 1,3 m do solo), volume, altura total, área basal, ou qualitativos – como aqueles usados para a representação do estado das árvores (dominante, dominada, viva, morta, com falhas) ao longo das medições realizadas periodicamente, dentre outros.

Salienta-se que as variáveis que representam tais dados podem ser classificadas segundo dois aspectos: quanto à representação dos seus valores (tipo do dado) e quanto à natureza da informação (tipo da variável). Tipo de dado diz respeito à forma em que um dado é armazenado, e os tipos de variáveis indicam a natureza da informação, ou como ela deve interpretada (GOLDSCHMIDT; PASSOS, 2005). Variáveis podem ser classificadas em:

> Nominais ou categóricas: utilizadas para nomear ou rotular elementos, podendo assumir valores pertencentes a um conjunto finito e pequeno de estados possíveis.
>
> Discretas: muito semelhantes às categóricas, mas os valores possíveis têm um ordenamento que possui algum significado.
>
> Contínuas: são variáveis quantitativas cujos valores possuem uma relação de ordem entre eles, e seu conjunto pode ser finito ou infinito.

3.4 Estatística descritiva

A estatística descritiva consiste na obtenção, análise e interpretação de dados numéricos através da criação de instrumentos adequados: quadros, gráficos e indicadores numéricos (REIS, 1996). Huot (2002) define estatística descritiva como o conjunto das técnicas e das regras que resumem a informação recolhida sobre uma amostra ou uma população, e isso sem distorção nem perda de informação. A estatística descritiva pode ser considerada como um conjunto de técnicas utilizadas para resumir o conjunto dos dados recolhidos numa dada investigação, que são organizados, geralmente, através de números, tabelas e gráficos.

Morais (2005) destaca que a estatística descritiva pretende proporcionar relatórios que apresentem informações sobre a tendência central e a dispersão dos dados. Para tal, deve-se evidenciar: valor mínimo, valor máximo, soma dos valores, contagens, média, moda, mediana, variância, e desvio padrão. A Figura 4 exemplifica o uso da estatística descritiva na mensuração florestal.

```
Sítio: 21
Idade:: 24
-----------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------
  volumeTotal     volumeTotalHA         B                S            N              NHa
 Min.   :0.5538   Min.   :17.72   Min.   : 4.670   Min.   :21   Min.   :24.00   Min.   : 889.7
 1st Qu.:0.7158   1st Qu.:25.48   1st Qu.: 5.985   1st Qu.:21   1st Qu.:28.00   1st Qu.: 969.6
 Median :0.9306   Median :30.74   Median : 6.484   Median :21   Median :30.00   Median :1001.7
 Mean   :1.0973   Mean   :37.62   Mean   : 7.260   Mean   :21   Mean   :30.26   Mean   :1025.7
 3rd Qu.:1.1946   3rd Qu.:42.54   3rd Qu.: 8.023   3rd Qu.:21   3rd Qu.:32.50   3rd Qu.:1060.3
 Max.   :2.4399   Max.   :87.98   Max.   :11.976   Max.   :21   Max.   :40.00   Max.   :1445.3
-----------------------------------------------------------------------------------------
Idade:: 36
  volumeTotal     volumeTotalHA          B                S            N              NHa
 Min.   :1.612   Min.   : 58.07   Min.   : 8.806   Min.   :21   Min.   :24.00   Min.   : 887.3
 1st Qu.:2.412   1st Qu.: 78.70   1st Qu.:11.104   1st Qu.:21   1st Qu.:28.00   1st Qu.: 969.6
 Median :2.708   Median : 91.35   Median :12.585   Median :21   Median :30.00   Median :1001.9
 Mean   :2.848   Mean   : 96.35   Mean   :12.766   Mean   :21   Mean   :30.44   Mean   :1027.3
 3rd Qu.:3.113   3rd Qu.:100.46   3rd Qu.:13.325   3rd Qu.:21   3rd Qu.:32.50   3rd Qu.:1062.4
 Max.   :4.892   Max.   :146.49   Max.   :17.728   Max.   :21   Max.   :43.00   Max.   :1445.3
-----------------------------------------------------------------------------------------
Idade:: 48
  volumeTotal     volumeTotalHA         B                S            N              NHa
 Min.   :2.821   Min.   :103.9   Min.   :12.03   Min.   :21   Min.   :24.00   Min.   : 887.3
 1st Qu.:3.910   1st Qu.:130.9   1st Qu.:15.52   1st Qu.:21   1st Qu.:28.00   1st Qu.: 970.6
 Median :4.246   Median :145.7   Median :16.93   Median :21   Median :30.00   Median :1006.1
 Mean   :4.406   Mean   :149.7   Mean   :16.96   Mean   :21   Mean   :30.35   Mean   :1029.7
 3rd Qu.:4.933   3rd Qu.:161.5   3rd Qu.:18.50   3rd Qu.:21   3rd Qu.:32.25   3rd Qu.:1070.1
 Max.   :6.734   Max.   :209.6   Max.   :21.30   Max.   :21   Max.   :43.00   Max.   :1445.3
-----------------------------------------------------------------------------------------
```

Figura 4. Apresentação de dados.

A Figura 4 apresenta os dados por parcela, no caso apresentado temos um determinado sítio e dados em diferentes idades, através disso um analista florestal pode identificar inconsistências existentes entre as variáveis.

3.5 Modelagem matemática

Aguirre (2007) destaca a modelagem matemática como a área do conhecimento que estuda maneiras de desenvolver e implementar modelos matemáticos de sistemas reais. A mesma tem como objetivo final fornecer uma descrição matemática de um dado fenômeno do mundo real. Tal descrição, que geralmente é feita por meio de equações, é chamada de modelo matemático. Neste trabalho destacam-se os modelos de regressão, e os modelos de crescimento e produção.

Dentre os modelos matemáticos, destacam-se os de predição ou projeção, o primeiro tem como finalidade inferir sobre condições ou comportamentos atuais baseando-se em informações existentes (histórico) relacionando variáveis disponíveis em um conjunto de dados, em suma são utilizados quando se deseja saber alguma informação futura ou uma estimativa do que acontecerá futuramente. Segundo Assunção (2012) modelos preditivos têm sido utilizados cada vez mais pelo mercado a fim de auxiliarem as empresas na mitigação de riscos, expansão de carteiras, retenção de clientes, pretensão de fraudes, entre outros objetivos. No que se refere a qualidade de dados em modelos preditivos, o autor afirma que durante o desenvolvimento desses modelos é comum, dentre as variáveis preditas, algumas possuírem dados não preenchidos (*missing data*) sendo necessário assim adotar algum procedimento para tratamento dessas variáveis. Com base nisso, a existência de dados não preenchidos é um problema relacionado a qualidade dos dados, podendo assim causar prejuízo ao usuário do modelo ao obter predições errôneas.

Diversas são as técnicas utilizadas diretamente em modelos de predição a fim de entender as relações existentes entre os dados presentes no modelo, e através disso formular conclusões baseando-se no resultado dessas técnicas (predição). Uma delas é a mineração de dados, que permite a definição de mapeamentos e regras capazes de revelar conhecimento novo sobre um domínio. Além disso, a mineração de dados aborda técnicas como modelos de regressão, que segundo Maita (2016) consistem em buscar uma função que represente, de forma aproximada, o comportamento apresentado pelo fenômeno em estudo. Outra técnica conhecida para predição são as Redes Neurais Artificiais (RNA). Maita (2016) diz que as RNA consistem em buscar similaridades entre dados para definir um conjunto finito de grupos de dados, organizando os dados em agrupamentos de forma a maximizar a

similaridade intra-agrupamentos e minimizar a similaridade interagrupamentos, além de consistir em determinar um conjunto não conhecido de classes para dados de forma também a descrevê-los. Cabe ainda destacar as máquinas de vetores de suporte, que segundo Maita (2016) são utilizados quando envolve uma componente temporal, isto é, representa problemas em que o interesse é prever o valor futuro de uma determinada variável ou em valores anteriores da variável de interesse e de outras variáveis (multi-variável).

3.5.1 Modelo de regressão

Montgomery (2001) define regressão como "uma técnica estatística para modelar e investigar a relação entre variáveis". Weisberg (2005) defende que a regressão, assim como a maioria das técnicas estatísticas, tem como meta sumarizar os dados observados de forma mais simples, útil e elegante possível. Para Gujarati (2000), a análise de regressão estuda a dependência da característica de interesse, em uma relação a uma ou mais variáveis, com objetivo de estimar o valor médio da característica de interesse em função dos valores conhecidos das variáveis de controle.

Modelos de regressão são modelos matemáticos que relacionam o comportamento de uma variável Y com outra X. Quando a função f que relaciona duas variáveis é do tipo $f(X) = a + b X$ temos o modelo de regressão simples. A variável X é a variável independente da equação enquanto $Y = f(X)$ é a variável dependente das variações de X. O modelo de regressão é chamado de simples quando envolve uma relação causal entre duas variáveis. O modelo de regressão é multivalorado quando envolve uma relação causal com mais de duas variáveis. Isto é, quando o comportamento de Y é explicado por mais de uma variável independe $X_1, X_2, ...X_n$.

Soares (2010) afirma que a análise de regressão é uma das ferramentas mais utilizadas para a análise de dados, sendo aplicada em quase todas as áreas de conhecimento, incluindo: engenharia, física, química, administração, ciências biológicas e agronomia. A simplicidade e utilidade dessa técnica são resultantes do conceito lógico de utilizar uma equação matemática para expressar a relação entre variáveis, sendo interessante devido à base matemática e bem desenvolvida teoria estatística (MONTGOMERY, 2001).

3.5.1.1 Aplicação de modelos regressão: como o dado irá trabalhar na estimação de valores

A aplicação mais comum dos modelos de regressão é na estimativa da característica de interesse *Y* para um conjunto de valores das variáveis de controle do modelo. Soares (2010) apresenta as formas de se obter tais estimativas:

1) Estimação da resposta média
2) Predição

A diferença entre essas duas observações recai no fato de que a estimativa da resposta média é calculada com base em um conjunto de valores das variáveis de controle que foi utilizado para estimar o modelo de regressão, enquanto a previsão de novos valores é estimada através de um conjunto de valores para as variáveis de controle.

3.5.2 Modelos de crescimento e produção

Modelos de crescimento e produção florestal são de grande importância nos estudos realizados em Ciências Florestais, pois permitem simular a dinâmica natural do povoamento, fornecendo resultados fundamentais para uma série de tarefas dentro da empresa florestal, como: o planejamento estratégico da empresa, o planejamento da colheita, a aferição de resultados econômicos, dentre outros. O modelo de crescimento e produção mais utilizado pelas empresas brasileiras até 2015 foi o de Clutter (Romaniuk, 2015). Tal modelo é um sistema composto por dois modelos, um para estimar a área basal futura e outro para estimar o volume futuro, e assim, podendo ser utilizado para simular o efeito de diferentes regimes de desbaste (processo de redução controlada e acelerada do número de árvores).

Modelos de crescimento e produção são constituídos por sistemas de equações, que por sua vez envolvem um grande número de funções, nas quais as variáveis dependentes de uma função, em alguns casos referem-se às variáveis preditas em outra. Tal característica induz a erros no processo de modelagem em consequência da propagação de erros de uma função para outra (GERTNER, 1991).

Há diversas classificações para modelos de crescimento e produção, dentre as quais destacam-se os modelos empíricos: modelos de em nível de povoamento (MNP), modelos de distribuição de diâmetros (MDD) e modelos de árvores individuais (MAI) (Vanclay,1994). Tais modelos são representados por equações nas quais o

resultado é válido apenas para as condições específicas em que foram estabelecidos, sendo altamente dependentes dos dados utilizados para seu ajuste (CAMPOS e LEITE, 2013). A Figura 5 exemplifica os modelos de crescimento e produção e seus tipos.

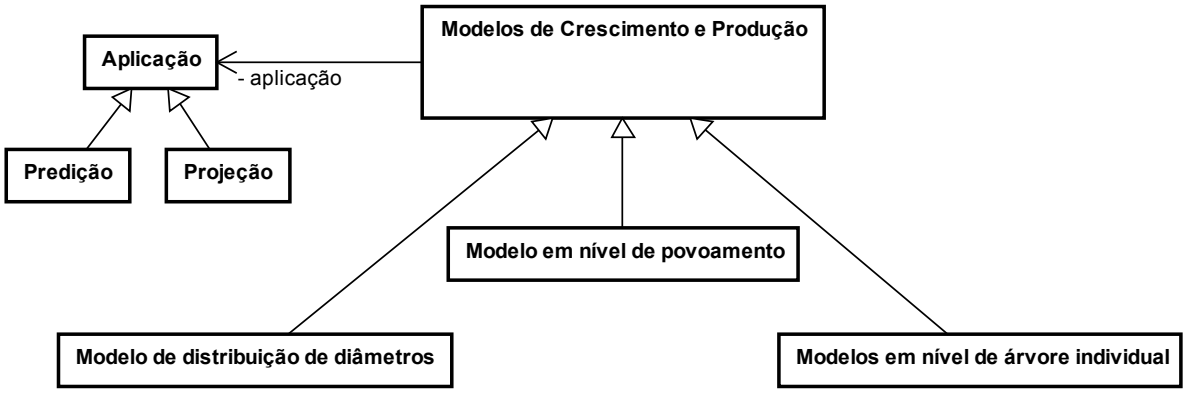


Figura 5. Tipos de Modelos de Crescimento e Produção.

Fonte: Adaptado de Campos e Leite (2013)

A Figura 5 apresenta uma relação como os modelos de crescimento e produção podem ser apresentados, primeiramente eles podem ser aplicados como modelos de predição ou projeção, como citado anteriormente, modelos de predição tem como intuito prever comportamentos futuros através de um histórico, e os de projeção visam prever comportamentos através de uma variável atual.

Na área florestal os modelos de crescimento e produção podem ser separados em modelos de árvore individual onde se analisa a árvore como variável, um conceito de análise de um elemento único, já os modelos de distribuições de diâmetro incorporam diversas árvores que estejam de terminada classe de diâmetro, por exemplo a classe de diâmetro 2 possui todas as árvores que tem diâmetro maior ou igual a dois e menor que 3. Por fim, nos modelos em nível de povoamento ou de povoamento total são analisadas áreas que contem centenas de árvores, tendo como objeto de análise um local ou sítio que representam um povoamento florestal.

3.5.2.1 Modelo em nível de povoamento ou povoamento total (MNP)

Modelos de povoamento total são aqueles que consideram a produção por unidade de área. Esses modelos são tipicamente inflexíveis para avaliar multiprodutos e grande número de tratamentos, embora sejam limitados a predizer apenas a produção total (BURKHART, 1977). De acordo com Campos e Leite (2013), tais

modelos estimam o crescimento e, ou, a produção a partir de atributos em nível de povoamento, como idade, área basal, e índice local. Sendo Y a variável dependente, eles podem ser subdivididos em (CAMPOS; LEITE, 2013):

a) Y = f(I);

b) Y = f(I, S); e

c) Y = f(I, S ou Hd, B, N ou q),

Em que: I é a idade do povoamento; S é o Índice de local; Hd é a altura dominante; B é a Área basal; N é o número de árvores por hectare; e q é o diâmetro médio.

Espera-se que, ao incluir mais variáveis independentes ao modelo, estimativas mais consistentes possam ser obtidas. Alguns destes modelos são apresentados na Tabela 1.

Tabela 1. Principais modelos em nível de povoamento.

Relação Funcional	Modelo
$V = e^{\beta_0+\beta_1(\frac{1}{I})} + \varepsilon$	Schumacher
$V = \frac{\beta_0}{1+\beta_1 e^{-\beta_2 I}} + \varepsilon$	Logístico
$V = \beta_0 + e^{\beta_1-\beta_2 I} + \varepsilon$	Gompertz
$V = \frac{\beta_0}{(1+\beta_1 e^{-\beta_2 I})^{\frac{1}{\beta_3}}} + \varepsilon$	Richards
$lnB_2 = LnB_1\left(\frac{I_1}{I_2}\right) + \alpha_0\left(1-\frac{I_1}{I_2}\right) + \alpha_1\left(1-\frac{I_1}{I_2}\right)S + \varepsilon$ $LnV_2 = \beta_0 + \frac{\beta_1}{I_2} + \beta_2 S + B_3 LnB_2 + \varepsilon$	Clutter (1963)

Fonte: Campos e Leite (2013)

B2 é a área basal do povoamento na idade futura (m² ha-1); B1 é a área basal do povoamento na idade atual; I1 é a idade atual, em meses; I2, a idade futura; S é o índice de local; V2 é o volume de madeira com casca do povoamento, na idade futura (m³ ha-1); com $\alpha 1$ e $\alpha 2$, que são os coeficientes do modelo de área basal; $\beta 1 ... \beta 4$ são os coeficientes do modelo volumétrico; e ε é o erro aleatório.

3.5.2.2 Modelo de distribuição de diâmetros (MDD)

Modelos de distribuição de diâmetros permitem estimar a produção por classe de diâmetro, possibilitando uma análise de produtividade em relação às dimensões das árvores do povoamento (FRAGA FILHO, 2016). O modelo de distribuição de diâmetros tem como objetivo fornecer informações sobre a estrutura do povoamento, permitindo assim a estimativa da produção por classes de diâmetro. Com tais informações é possível realizar o planejamento do corte, de desbastes, os tratamentos silviculturais, estimar multiprodutos como, por exemplo, madeira para serraria, celulose e energia (LEITE, 1990).

Neste tipo de modelo os parâmetros de uma equação ou de um sistema de equações são correlacionados com atributos do povoamento na idade atual, tais como idade, altura, diâmetro, área basal, mortalidade e ingresso, de acordo com o nível pretendido. Na projeção os parâmetros são estimados em uma idade futura em função dos mesmos atributos e dos parâmetros obtidos em uma idade atual (CAMPOS; LEITE, 2013). A principal característica dos MDD é o emprego de uma função de densidade de probabilidade, sendo comum o uso da função Weibull, uma das mais utilizadas nessa modelagem, especialmente pela sua grande flexibilidade (CAMPOS; LEITE, 2013). Tal flexibilidade pode ser observada na Figura 6.

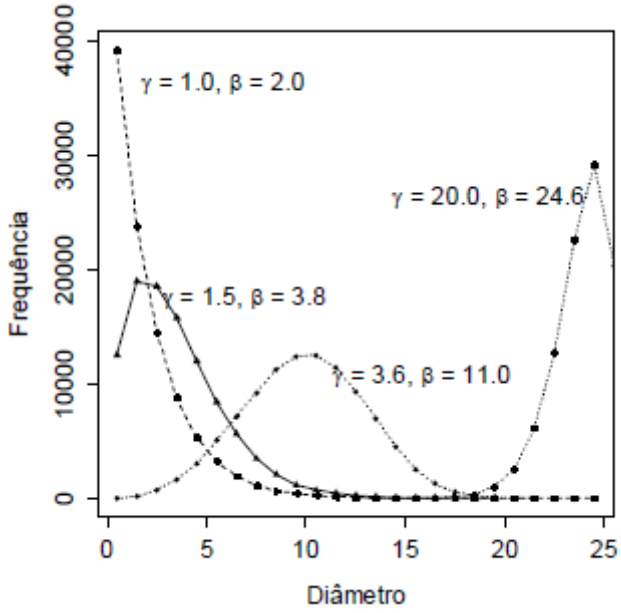


Figura 6. Função de Densidade de Probabilidade Weibull, para diferentes valores de ɣ (forma) e β (escala).

Fonte: Adaptado de Campos e Leite (1990)

3.5.2.3 Modelo em nível de árvores ou de árvore individual (MAI)

Modelo em nível de árvores ou de árvore individual (MAI) são sistemas de equações e procedimentos usados para estimar crescimento e produção de povoamentos, utilizando árvores como unidade de modelagem, em vez de parcela (CAMPOS e LEITE, 2013). Difere do modelo de povoamento total pela necessidade de especificar o tamanho de cada árvore dentro do povoamento, em relação ao seu crescimento em diâmetro, altura, e, em alguns casos, do tamanho da copa e sua posição no dossel aumentando a complexidade para definição dos dados e de sistemas para seu ajuste e aplicação (VANCLAY, 1994).

Além disso, Campos e Leite (2013) ressaltam que esse tipo de modelo gera resultados detalhados das estruturas horizontal e vertical e pode ser empregado para simular tratamentos silviculturais como o desbaste e a desrama artificial. Fraga Filho (2016) apresenta um fluxograma representando o procedimento necessário para obtenção de um modelo de árvore individual, o mesmo pode ser observado na Figura 7.

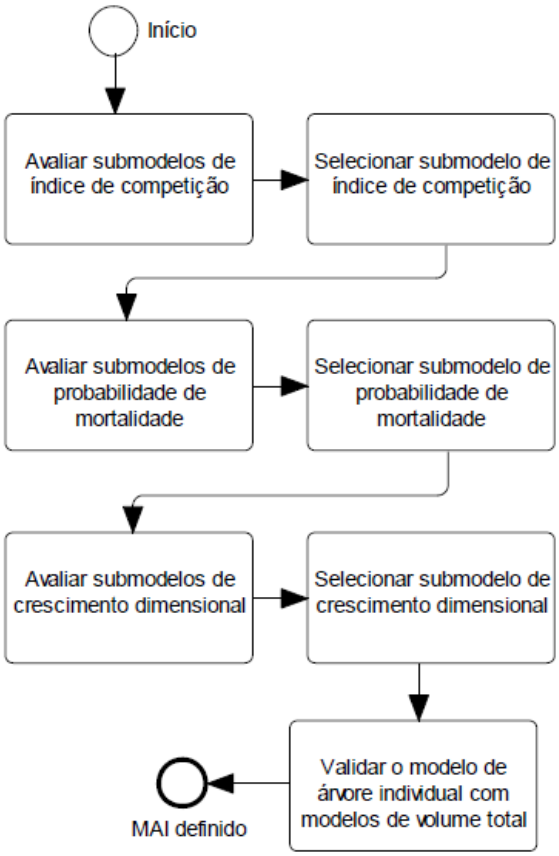


Figura 7. Determinação do modelo de árvores individuais.

Fonte: Adaptado de Fraga Filho (2016)

3.5.2.4 Dados para modelos de crescimento e produção

Uma vez compreendida a importância dos modelos de crescimento e produção para fins de planejamento e manejo florestal, é preciso entender a natureza dos dados necessários para sua construção e das várias fontes desses dados. Campos e Leite (2013) afirmam que tão importante quanto definir o modelo adequado para uma situação particular é conhecer características dos dados imprescindíveis para sua contribuição, além do modo de realizar a amostragem e suas consequências na eficiência do modelo. A Figura 8 apresenta as diferentes formas de representação dos dados em modelos de crescimento e produção.

(A)

Parcela	M	ARV	I	DAP	HT	V/Ha
1	2003	1	1	12.3	15.1	10.8
1	2004	1	2	13.2	17.2	12.1
1	2005	1	3	14.1	19.1	11.9
1	2003	2	1	8.1	12	9.2
1	2004	2	2	9	10	11
1	2005	2	3	11	12	13.6
2	2004	1	1	9.5	11	9.8
2	2005	1	2	12.1	12	12.1
2	2006	1	3	9.8	14.1	11.6
2	2004	2	1	8.6	10.9	11.3
2	2005	2	2	7	9.8	9.12
2	2006	2	3	12	13	13.9

(B)

Parcela	ARV	M1	M2	I1	I2	DAP1	DAP2	HT1	HT2	V1	V2
1	1	2003	2004	1	2	12.3	13.2	15.1	17.2	10.8	12.1
1	1	2004	2005	2	3	13.2	14.1	17.2	19.1	12.1	11.9
1	2	2003	2004	1	2	8.1	9	12	10	9.2	11
1	2	2004	2005	2	3	9	11	10	12	11	13.6
2	1	2004	2005	1	2	9.5	12.1	11	12	9.8	12.1
2	1	2005	2006	2	3	12.1	9.8	12	14.1	12.1	11.6
2	2	2004	2005	1	2	8.6	7	10.9	9.8	11.3	9.12
2	2	2005	2006	2	3	7	12	9.8	13	9.12	13.9

(C)

Parcela	M1	M2	I1	I2	Classe DAP	N1	Prob1	V1/Classe	N2	Prob2	V2/Classe	V1	V2
1	2003	2004	1	2	9	1	0.5	10	1	0.5	11.55	20	23.1
1	2003	2004	1	2	11	0	0	0	0	0	0	20	23.1
1	2003	2004	1	2	13	1	0.5	10	1	0.5	11.55	20	23.1
1	2003	2004	1	2	15	0	0	0	0	0	0	20	23.1
1	2004	2005	2	3	9	1	0.5	11.55	0	0	0	23.1	25.5
1	2004	2005	2	3	11	0	0	0	1	0.5	12.75	23.1	25.5
1	2004	2005	2	3	13	1	0.5	11.55	0	0	0	23.1	25.5
1	2004	2005	2	3	15	0	0	0	1	0.5	12.75	23.1	25.5
...													

(D)

Parcela	M1	M2	I1	I2	V1	V2
1	2003	2004	1	2	20	23.1
1	2004	2005	2	3	23.1	25.5
2	2004	2005	1	2	21.1	21.22
2	2005	2006	2	3	21.22	25.5

Figura 8. Origem dos dados de crescimento e produção.

Fonte: Fraga Filho (2016)

A Figura 8 apresenta formas que os dados podem ser representados e como eles evoluem a partir de cada modelo, em (A) temos o dado em sua forma mais básica, que são os dados para inventário florestal, os mesmos estão dispostos em variáveis como parcela, medição (M), árvore (ARV), idade (I), diâmetro (DAP), altura (HT), volume (V/Ha). Após isso, os dados em modelos de crescimento em produção, os mesmos que antes estavam representados para inventário florestal agora compõem

dados utilizados por modelos, algumas variáveis continuam já outras são adicionadas adaptando-se assim ao tipo do modelo, por exemplo, nos MAI e MNP apresentados em (B) e (D) respectivamente são adicionadas segundas variáveis como: medição, idade, diâmetro, altura e volume (M2, I2, DAP2, HT2, V2 respectivamente).

Por fim, nos modelos de distribuição diamétrica (MDD) apresentados em (C) são inseridos os conceitos de classe por diâmetro e surgem novas variáveis como: classe de diâmetros (classe DAP), número de árvores na classe na idade 1 (N1), número de árvores na classe na idade 2 (N2), Prob1 e Prob2 são probabilidades associadas ao total de árvores em determinada idade (I1 ou I2), o total é obtido pela multiplicação de Prob1 ou Prob2 pelo número de árvores na classe, e por fim V1/classeV2/classe são os volumes por classe de diâmetro.

Uma vez que a árvore é a unidade básica a ser medida, para melhorar o entendimento da quantificação de características qualitativas, considera-se definir estados atuais e estados associados, que podem ocorrer ao longo de um ciclo de corte, apresentados na Figura 9.

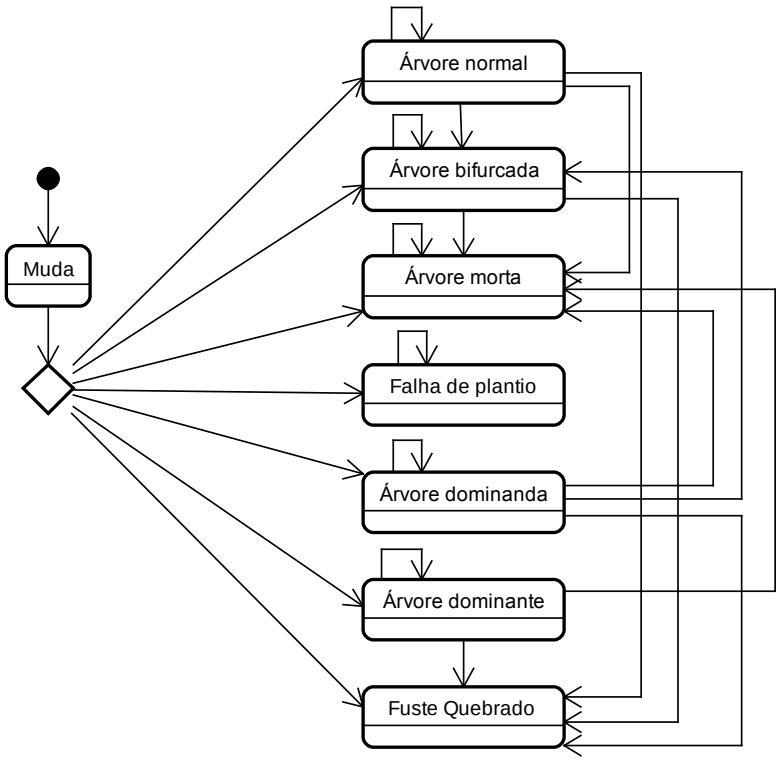


Figura 9. Estados associados a covas, árvores e fustes em parcelas permanentes.

Fonte: Adaptado de Campos e Leite (2013).

O diagrama representado na Figura 9 apresenta os estados, representados pelos retângulos, sendo o círculo o início da representação. As setas representam a

transição de um estado inicial para um estado associado desde o plantio até a primeira rotação de parcelas permanentes. Observa-se que, por exemplo, uma muda pode passar para todos os demais estados associados, exceto permanecer como muda. Assim como uma cova morta não realiza qualquer transição de estado a não ser classificada como morta pelos sucessivos períodos de remedição, diferente de uma cova normal que pode continuar normal, ou apresentar bifurcação, morrer ou ter um fuste quebrado. Salienta-se que estes estados são mantidos no registro do inventário florestal e representam o estado “cadastral” da cova (elemento comumente utilizado como identificador do local de plantio de cada árvore nas respectivas parcelas), não o estado real da cova na parcela.

A qualidade das estimativas e resultados obtidos por modelos de crescimento e produção nos diferentes níveis depende de muitos fatores, sendo a qualidade dos dados utilizados no ajuste o principal deles (MAGALHÃES, 2014). Uma vez que os dados que irão compor o modelo atendam expectativas relacionadas a qualidade de dados, espera-se obter modelos mais precisos e confiáveis para seus fins, dessa forma, a necessidade de tratamento dos dados para modelos de crescimento e produção torna-se fundamental no processo de planejamento e manejo florestal.

3.6 Problemas relacionados à Qualidade de Dados

Dentre os problemas relacionados a qualidade de dados na área florestal, destacam-se os relacionados à pontualidade, completude e consistência, demonstrando o que pode acarretar com a falta desses requisitos. A pontualidade é uma característica fundamental, Fraga Filho (2016) afirma que se espera que o dado esteja disponível no momento em que seja necessário realizar um estudo de prognose para avaliação de disponibilidade de múltiplos produtos em uma idade futura. Caso ocorra a indisponibilidade, os clientes que demandam a informação podem ficar frustrados causando estresse no ambiente organizacional e até mesmo levando membros da equipe a criarem dados irreais, ou usar métodos não recomendados para sua geração.

Tanto a completude quanto a consistência devem ser avaliadas em cada medição e entre medições. A falta de informações sobre o DAP de uma ou mais árvores em algumas parcelas pode prejudicar o ajuste de um modelo que seria promissor para determinado propósito (FRAGA FILHO, 2016).

3.7 Detecção e Correção de Erros

Localização e correção de erros são úteis toda vez que os dados foram coletados a partir de fontes passíveis de erros (por exemplo, aqueles em que a entrada manual foi realizada) ou adquiridos a partir de fontes cuja confiabilidade não é totalmente garantida e sofre pressão, por exemplo, para cumprimento de prazos, metas e outros (BATINI; SCANNAPIECA, 2006). É importante ressaltar que não é possível corrigir todos os erros, mesmo porque alguns deles não podem ser detectados, a ideia é detectá-los através de métodos de garantida de qualidade e informar o usuário sobre o mesmo, e a partir de aí buscar soluções.

Após a localização de registros com erros, a fim de corrigir erros, pode-se executar sobre eles a aquisição de novos dados. Porém sabe-se que este tipo de atividade é geralmente muito custoso, e, em todos os contextos em que são recolhidos dados para fins estatísticos, o uso de edições é geralmente preferido para corrigir dados errados.

O problema de localização de erros por meio de edições de campos de forma errada é normalmente referido como o problema de edição de imputação. Batini e Scannapieca (2006) apresentam um modelo teórico para estes problemas, sendo seus principais objetivos: Os dados em cada registro deve satisfazer todas as edições mudando o menor número de campos possível. Este é o chamado princípio da mudança mínima.

Quando a entrada de dados é necessária, deseja-se manter a distribuição de frequência conjunta de valores nos diferentes campos entre diferentes registros. Sendo assim, ao modificar o registro, por exemplo o valor de DAP, é importante observar se tal modificação pode alterar o comportamento das variáveis relacionadas a ele no registro e também no conjunto representado por outros registros. Caso um DAP numa idade inicial seja modificado para um valor comum apenas em idades perto do corte, o comportamento global de árvores naquela idade pode ser modificado. Erros cometidos nas medições de diâmetros e alturas das árvores são devido a erros do operador, defeitos no instrumento e condições de observação (AUGUSTO et al., 2015).

O desenvolvimento em idades futuras de florestas prognosticado por modelos de crescimento sempre inclui erros, cuja quantidade depende (i) variação aleatória no crescimento, (ii) descrição imprecisa do atual estado da floresta e (iii) os parâmetros

de modelos de crescimento e sua estimativa erros. Além disso, pode haver incerteza sobre a especificação incorreta de um modelo, seja pela má formulação, número e tipo de variáveis escolhidas e seu efeito, as quais podem introduzir erros sistemáticos (PIETILÄ et al., 2010).

Uma vez que o erro foi detectado e diagnosticado, surge a questão: como corrigi-lo? As correções vão de correção manual, exibindo o erro para o usuário e solicitando para o mesmo efetuar a correção redigitando o dado, outro processo de correção seria a eliminação do dado, seja ele um registro ou uma parcela e por fim como última alternativa podemos aplicar o submodelo proposto por Pienaar e Shiver (1981) representado pela equação.

$$Y_2 = Y_1 . e^{-\beta_0 (I_2{}^{\beta_1} - (I_1{}^{\beta_2})} + \varepsilon$$

Em que: Y2 é o diâmetro (em cm) ou a altura total (m) em idade futura; Y1 é o diâmetro (cm) ou altura (m) em idade atual; I2 é a idade futura (meses); I1 é a idade atual (meses), βi são os coeficientes do modelo e ε é o erro aleatório. De acordo com Fraga Filho (2016) o submodelo apresentado é utilizado para estimar o crescimento em diâmetro (DAP) e altura (HT), tais medidas podem apresentar erros do tipo DAP2 < DAP1 (diâmetro na idade 2 é menor que na idade 1), HT2 < HT1 (altura na idade 2 é menor que altura na idade 1).

O submodelo que Pienaar e Shiver (1981) ajusta os dados através de estimativas. Em caso de dado faltante como exemplificado acima, remove-se a medição faltante e estima-se seu valor com os dados da parcela no qual ele pertence utilizando do modelo proposto por Pienaar e Shiver (1981), servindo assim como mais uma maneira de correção dos dados das bases.

3.7.1 Dados longitudinais

Pesquisas longitudinais são aquelas em que as mesmas (ou diferentes) variáveis são medidas para os mesmos indivíduos em, pelo menos, dois pontos distintos do tempo, em nosso caso, os indivíduos são as árvores, e as variáveis associadas a elas estão presentes em bases dados que armazenam todo tipo de característica pertinente a árvore.

A análise dos dados longitudinais permite estudar a mudança de comportamento, por exemplo, de um mesmo indivíduo ou variações entre indivíduos ao longo do tempo. Os resultados provenientes de análises desse tipo em estudos de

cunho social ou ambiental auxiliam as organizações competentes a tomarem decisões corretas para solucionar seus problemas (CARNEIRO, 2012). No que diz respeito a dados longitudinais, medidos ao longo do tempo, dois tipos de incompletudes podem ser identificados: dados truncados e dados censurados (BATINI; SCANNAPIECA, 2006).

Dados truncados correspondem a observações que são descartados a partir do conjunto de dados analisados. Por exemplo, parcelas abaixo de 24 meses ou árvores com DAP abaixo de 5 cm. Já os dados censurados correspondem aos dados que sabemos com certeza que não foram medidos antes de um período *p1* (dados censurados à esquerda) ou depois de um período *p2* (dados censurados à direita). Como um exemplo de dados censurados à esquerda, suponha que estamos interessados em realizar a classificação de sítio de um povoamento; poderíamos ter apenas dados disponíveis depois de uma medição, dados de medições anteriores poderiam não estar disponíveis. A Figura 10 ilustra os casos mencionados.

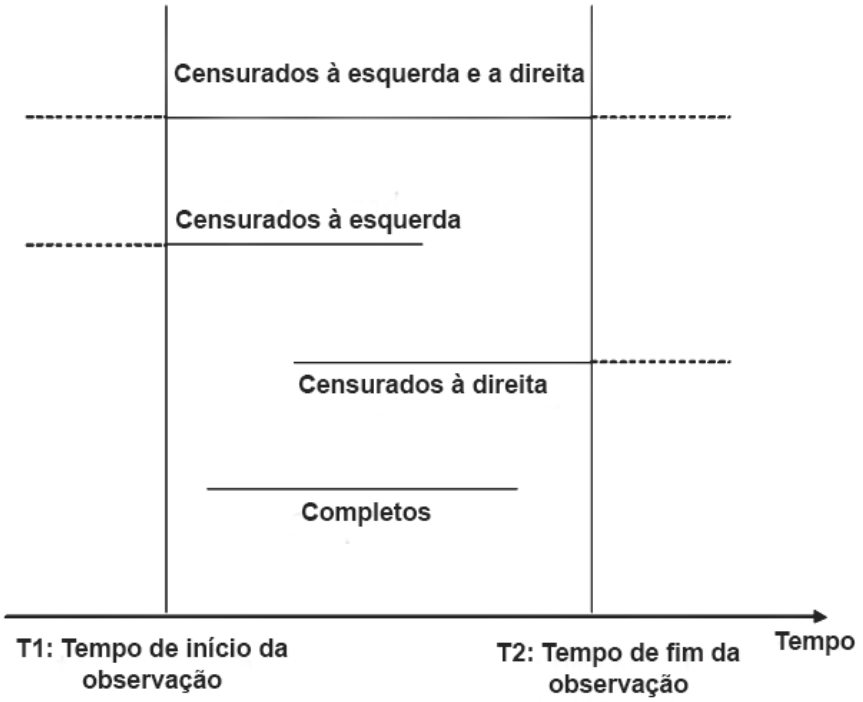


Figura 10. Tipos de dados incompletos.

Fonte: Adaptado de Batini e Scannapieca (2006)

Para a detecção deste tipo de problema histogramas e distribuições de freqüência podem ser utilizados por um especialista no domínio. A descoberta de *outliers* consiste em identificar valores altos ou baixos demais em relação à variável

analisada. Tipicamente, um *outlier* é atribuível a uma das seguintes causas na medição de dados:

1. O objeto não foi corretamente observado na gravação dos dados ou durante o registro no banco de dados.
2. Trata-se de uma população diferente, em relação a outros valores, mesmo considerando diferentes períodos de medições.
3. Que é correto, mas representa um evento raro.

Em todos os casos é necessário analisar a fim de identificar a origem do desvio. No caso 2 pode ocorrer de uma mesma parcela permanente ser medida em 2 inventários por equipes diferentes e em um dos casos a equipe inverteu a ordem de caminhamento na parcela.

3.8 Procedimentos de consistência

Para demonstrar a avaliação de características de qualidade de dados florestais, pode-se realizar operações de consistência a fim de explorar a base de dados em busca de inconformidades, para cada tipo de base seja ela de um IFC ou de modelos de crescimento (MNP, MAI, MDD) existem procedimentos específicos a serem realizados. Fraga Filho (2016) afirma que tais operações de consistência têm como intuito minimizar a inclusão de erros no processo de modelagem de crescimento e produção.

3.8.1 Procedimentos para árvore individual

Os estados dos fustes (partes da árvore, do solo aos primeiros ramos, tronco) representados por variáveis categóricas em ambos os casos são apresentados na Figura 11.

A - Dominada
B - Bifurcada
D - Dominante
F - Falha
I - Inclinada
M - Morta
N - Árvore Normal
Q - Quebrada

Figura 11 – Estados dos fustes representados por uma variável categórica.

Fonte: Fraga Filho (2016)

Tomando como partida nos códigos acima apresentados, Fraga Filho (2016) destaca os procedimentos de consistência:

a) Identificação de árvores com DAP e HT = 0 e estado diferentes de M, F ou A.

Estado	Parcela	HT	Ano	Idade	DAP
D	10	37.3	2010	93	13.2
N	11	0	2011	50	0
D	12	40.2	2010	61	13.7
N	14	30.3	2012	72	14.2
D	15	17.4	2013	75	16.1

Quadro 1 Exemplo de árvore com DAP e HT=0 e estado diferentes de M, F, A.

Pode-se perceber que a segunda linha representa um exemplo de árvore que seria detectada por esse tipo de análise.

b) Identificar árvores com DAP2 < DAP1 entre medições consecutivas i e i+1, sendo i o ano da medição;

Estado	Parcela	Ano1	Ano2	DAP1	DAP2
D	10	2009	2010	14.3	13.2
N	11	2010	2011	15.1	16.1
D	12	2008	2010	14.3	15.5
N	14	2011	2012	14.6	17.2
D	15	2002	2013	15.5	18.2

Quadro 2 Exemplo de árvores com DAP2 < DAP1 entre medições consecutivas i e i+1, sendo i o ano da medição.

A linha 1 do quadro exemplifica o método proposto, onde o DAP2 (13.2) é menor que o DAP1(14.3) em medições consecutivas 2009,2010.

c) Identificação de número de fustes mortos (estado = M) ao longo das medições.

Estado	Árvore	Medição 1	Medição 2	Medição 4
M	1	16.1	19.2	20.3
M	2	15.8	0	19.7
M	3	15.2	10.4	19.3
M	4	15.9	17.8	19.8
M	5	15.3	17.7	22.1

Quadro 3 Exemplo de número de fustes mortos (estado = M) ao longo das medições.

Percebe-se que o campo código todo o registro tem seu estado = M, ou seja, elas continuaram a ser medidas com estado igual a Morta.

d) Identificação de clones (material genético), parcelas com mais de um clone em diferentes anos (medições).

Estado	Árvore	Medição 1	Medição 2	Medição 4
N	1	16.1	19.2	20.3
D	1	16.1	19.2	20.3
N	3	15.2	10.4	19.3
D	4	15.9	17.8	19.8
N	5	15.3	17.7	22.1

Quadro 4 Exemplo de número de clones em parcela.

O quadro demonstra um exemplo de clone (árvores com números iguais) na linha 1 e 2, porém torna-se necessário avaliar uma vez que o estado do clone é diferente do estado do original (N diferente de D).

e) Identificação do número de parcelas por clone;

Estado	Árvore	Medição 1	Medição 2	Parcela
N	1	16.1	19.2	1
D	1	16.1	19.2	0
N	3	15.2	10.4	2
D	4	15.9	17.8	3
N	5	15.3	17.7	4

Quadro 5 Exemplo de clone com parcela indefinida.

A última coluna apresenta a parcela respectiva do clone apresentado na segunda linha da coluna árvore, porém a parcela apresenta valor igual a 0 sendo que seu clone possui parcela igual a 1, torna-se necessário investigar o porquê desta ocorrência.

f) Identificação de parcelas sem espaçamento definido

Estado	Árvore	Medição 1	Espaçamento	Parcela
N	1	16.1	3x3	1
D	2	16.4	0	2
N	3	15.2	3x3	3
N	4	15.9	3x3	4
D	5	15.3	3x3	5

Quadro 6 Exemplo de parcela com espaçamento indefinido.

A coluna espaçamento indica o tamanho da parcela relativa, 3x3 significa a quantidade em metros da parcela, no caso 3 metros por 3 metros, uma parcela praticamente não existe sem esse dado, na linha 2 temos uma parcela que está sem espaçamento, portanto é um caso a ser analisado.

g) Identificação de parcelas com mais de um espaçamento em idades diferentes.

Estado	Árvore	Medição 1	Espaçamento	Parcela	Ano
N	1	16.1	3x3	1	2010
D	2	16.4	0	2	2012
N	3	15.2	5x5	1	2014
N	4	15.9	3x3	4	2015
D	5	15.3	3x3	5	2012

Quadro 7 Exemplo de parcela com espaçamento diferente.

O quadro acima apresenta a parcela 1 com espaçamento de 3x3 em 2010 e em 2014 a mesma apresenta um espaçamento 5x5, tal discrepância não pode existir, portanto, mais um caso a ser avaliado.3

h) Identificar árvores que não atendam IDADE2 – IDADE1=1

Estado	Parcela	Idade 1 (Mês)	Idade 2 (Mês)	DAP1	DAP2
D	10	36	48	14.3	13.2
N	11	24	36	15.1	16.1
D	12	12	36	14.3	15.5
N	14	12	24	14.6	17.2
D	15	12	24	15.5	18.2

Quadro 8 Exemplo de árvores que não atendam IDADE2 – IDADE1=1.

A linha 3 do quadro exemplifica o método proposto, onde a idade 1 e a idade 2 apresentam uma diferença de 24 meses, ao invés de 12.

Espera-se que por meio de tais procedimentos seja possível identificar não-conformidades na base de dados quanto à qualidade dos dados ali presente, e através disso tomar medidas para realizar a consistência da mesma.

3.8.2 Incremento diamétrico

Demolinari (2006) apresenta uma técnica que utiliza de gráficos de barras coloridas para mostrar a dinâmica da mudança das árvores pelas classes de diâmetro, em diferentes capacidades produtivas. O potencial produtivo ou capacidade produtiva de um local ou sítio quando associado a florestas de produção é definido como a capacidade de uma região em produzir madeira (SCOLFORO, 2006). Tal capacidade pode ser expressa de modo qualitativo (alta, média, baixa) ou utilizando algum índice quantitativo, como o índice de local, representado pela letra “S”. (CAMPOS; LEITE, 2013).

Esse Índice é a altura dominante média do povoamento em uma idade específica (idade-índice), sendo assim, considerada uma medida quantitativa de capacidade produtiva do lugar (CAMPOS; LEITE, 2013). Campos e Leite (2013) afirmam que a classificação de índices de local é um método prático e com bom desempenho para avaliar a qualidade de um local, pois todos os fatores ambientais são refletidos de maneira interativa no crescimento em altura, que pode ser relacionada com o volume.

Tendo em vista o processo de incremento diamétrico, o mesmo é iniciado na identificação de cada classe diamétrica na idade inicial, após isso é feito o acompanhamento do crescimento das árvores que estavam nessa classe e suas

possíveis migrações para outras classes de diâmetro, a Figura 12 apresenta um diagrama de sequência que exemplifica o processo.

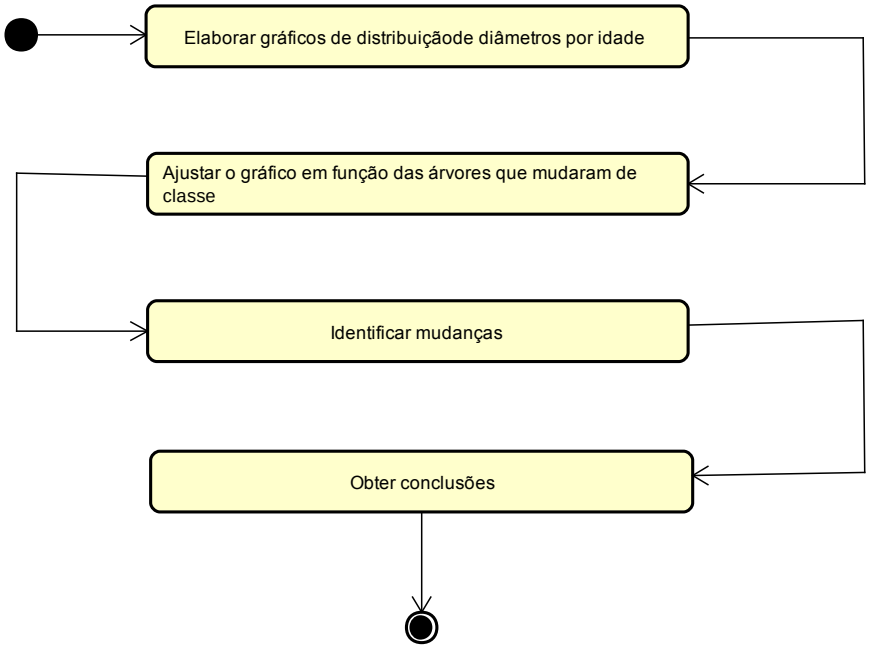


Figura 12 Processo de Incremento Diamétrico.

A Figura 12 representa o diagrama de atividades de todo o processo de incremento diamétrico proposto por Demolinari (2006), o início se dá na primeira atividade que é a construção dos gráficos de distribuição de diâmetro, tal gráfico tem a característica de ser colorido a fim de diferenciar as classes de diâmetro, após isso os mesmos gráficos são ajustados para uma idade futura apresentando através das cores definidas na idade inicial, quais classes sofreram alterações, após isso faz se a identificação de tais mudanças e se obtém as conclusões perante isso.

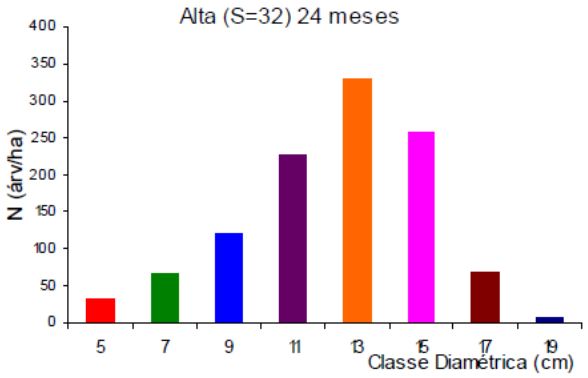


Figura 13 Distribuição de classe por produtividade igual 32 em 24 meses.

Fonte: Demolinari (2006)

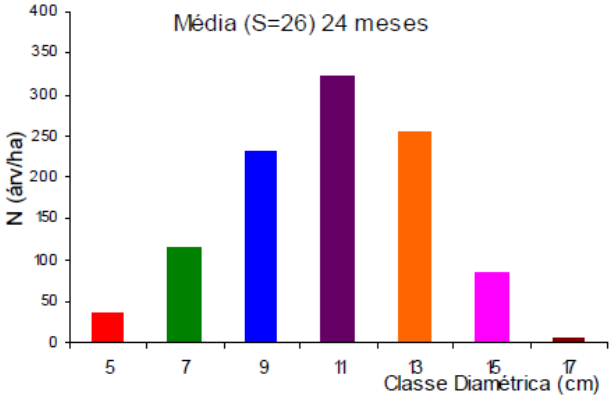


Figura 14 Distribuição de classe por produtividade igual 26 em 24 meses.

Fonte: Demolinari (2006)

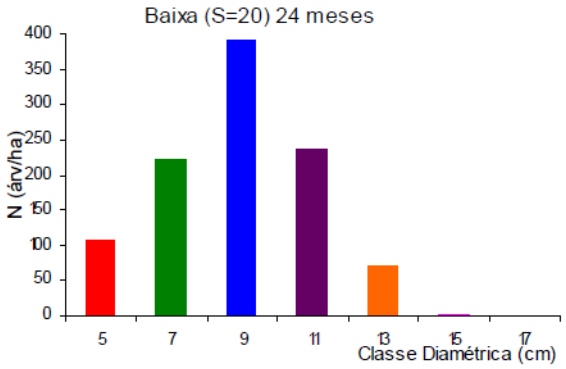


Figura 15 Distribuição de classe por produtividade igual 20 em 24 meses.

Fonte: Demolinari (2006)

As Figuras 13, 14, 15 representam a distribuição dos diâmetros aos 24 meses, para as classes de produtividade. Na Figura 13 temos capacidade produtiva alta, em

14 temos média e em 15 temos baixa. Isto significa que a Figura 13 representa o melhor local, ou seja, possui a melhor capacidade produtiva. Contudo, o mais interessante do incremento diamétrico é como ele apresenta a variação em relação ao tempo do crescimento de diâmetros. Demolinari (2006) apresenta a variação do diâmetro para 36 meses por meio das Figuras 16, 17, 18.

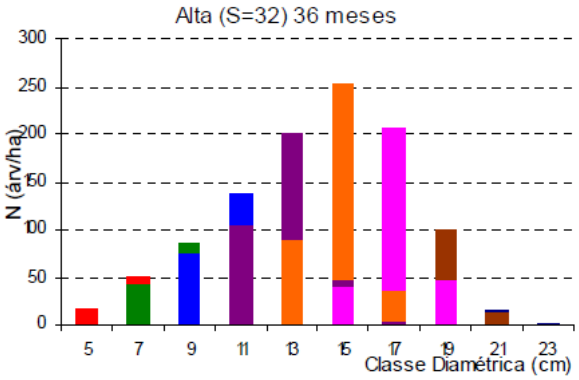


Figura 16 Gráfico de mudança de classe da área com produtividade igual a 32 em 36 meses.

Fonte: Demolinari (2006)

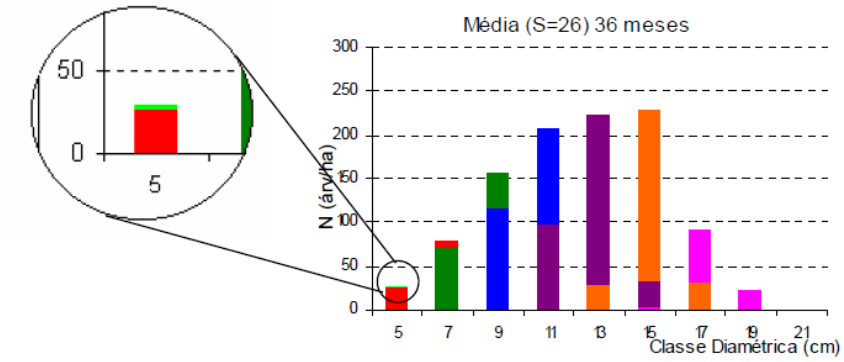


Figura 17 Gráfico de mudança de classe da área com produtividade igual a 26 em 36 meses.

Fonte: Demolinari (2006)

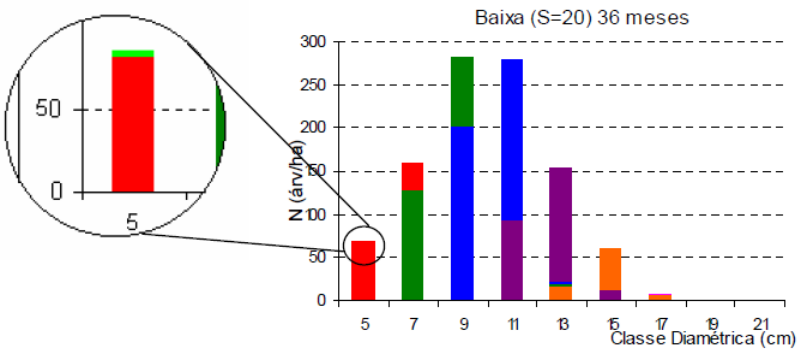


Figura 18 Gráfico de mudança de classe da área com produtividade igual a 20 em 36 meses.

Fonte: Demolinari (2006)

Observa-se que a distribuição é alta, média, e baixa para 16, 17, 18 respectivamente, isto indica que em 17 o processo de estagnação do crescimento das árvores menores ocorre mais cedo (DEMOLINARI, 2006), ou seja, onde há mais produtividade é o local onde houve menos árvores que mudaram de classe.

Com isso, por meio dessa variação de diâmetros existente pode-se elaborar processos de consistência a fim de encontrar dados anômalos, por exemplo, uma classe que não tinha crescimento considerável, teve um crescimento rápido em um curto período de tempo, pode-se então identificar problemas em medição ou algo correlacionado. A partir do método elaborado pelo autor pode-se retirar as seguintes verificações quanto a consistência dos dados apresentados:

1. Se árvores com idade inicial atingem tamanhos maiores em idades futuras.
2. Se árvores sobreviventes em um determinado período de tempo
3. Mortalidade em árvores que não mudaram de classe

Demolinari (2006) afirma que além das discrepâncias relevadas, pode-se identificar que variáveis de povoamento possuem relação direta com a capacidade produtiva, com diferentes taxas de crescimento ao longo do tempo, ou seja, que existe relação direta entre o tamanho das árvores e a dinâmica de crescimento. Árvores maiores em uma idade inicial atingem tamanhos maiores em idades futuras. Consequentemente, a probabilidade de as árvores morrerem é maior nas menores classes de diâmetro.

3.9 Métricas de qualidade de dados de bases florestais

Definir métricas para qualidade de dados tem como intuito a avaliação do estado atual e da evolução dos tratamentos de qualidade dos dados. Barbieri (2013) afirma que as métricas, como todas as medidas definidas em processos de qualidade deverão:

1. Ser atreladas a objetivos bem definidos.
2. Responder a questões associadas a esses objetivos.
3. Ser medidas definidas com clareza, que apontem elementos quantificáveis associáveis a objetivos de negócios, com formulações claras (como medir), valores definidos para análise (como analisar), com faixas aceitáveis e não aceitáveis (como interpretar), plano de ações no caso de discrepâncias, frequência de medição (quando medir), entre outros.

É no processo de avaliação dessas dimensões de qualidade de dados (consistência, completude, pontualidade, entre outras) que entram as métricas: regras de avaliação da qualidade para cada dimensão, sejam de aferição subjetiva ou objetiva. A aderência a uma lista de valores permitidos é algo que se pode medir objetivamente, neste caso a integridade da informação. Uma vez definidas as métricas, elas passam a ser aplicadas regularmente, constituindo-se um termômetro imprescindível para monitorar a qualidade dos dados da empresa e o resultado das ações de melhoria adotadas ao longo do tempo, auxiliando dessa forma na avaliação da consistência da base de dados.

Um exemplo de métrica seria a porcentagem de erros encontrados nas bases de dados florestais, a mesma pode ser representada pela seguinte expressão:

$$Porcentagem\ de\ Erro = \frac{Número\ de\ Erros}{Total} * 100$$

Sendo que o número de erros é a quantidade de erros contidos na base de dados florestais, o total representa a quantia de informações que foram analisadas.
Podemos citar outros exemplos de métricas baseados nos procedimentos de consistências apresentados até aqui:

A primeira que podemos definir para avaliação é a quantidade de DAP1's < DAP2's, e pode ser expressada pela expressão.

$$\text{Quantidade de DAP's com erro} = \frac{Quantidade\ de\ DAP's\ com\ erro}{Total\ de\ DAP's\ analisados}$$

Para as demais análises que buscam erros dentre os dados podemos definir métricas similares a acima, por outro lado, podemos apresentar métricas que totalizando a quantidade de registros anômalos, ou seja, que tenham dados que fogem dos métodos apresentados anteriormente, por exemplo parcelas clonadas, ou parcelas sem espaçamento definido, até mesmo identificar árvores que não atendam IDADE2 – IDADE1=1, quando a idade for representada em anos, na base de dados. O processo de métricas nos dá uma visão quantitativa de quão afetada está a qualidade de dados da base, dessa forma pode-se medir o quanto a qualidade está melhorando por exemplo, quando uma medida dessas está diminuindo com o passar do tempo.

3.10 Repositório de Pacotes do R (CRAN)

A linguagem R (R DEVELOPMENT CORE TEAM, 2005) possui um repositório de pacotes chamado CRAN ("Comprehensive R Archive Network"), o mesmo é a rede de distribuição do R, ou seja, uma coleção de sites onde estão armazenadas versões de códigos e documentações para o R que são idênticos no mundo todo.

De acordo com CRAN (2016) o repositório possui 9525 pacotes disponíveis atualmente, assim podemos buscar esses pacotes e sua documentação no CRAN de acordo com a necessidade e avanço na utilização do R. Tais pacotes são constantemente atualizados e podem ser facilmente acessados através de sites na internet, cobrindo uma ampla gama de variedade de programas estatísticos.

4 METODOLOGIA

Na Figura 19 é apresentado o fluxo das atividades realizadas na execução do trabalho.

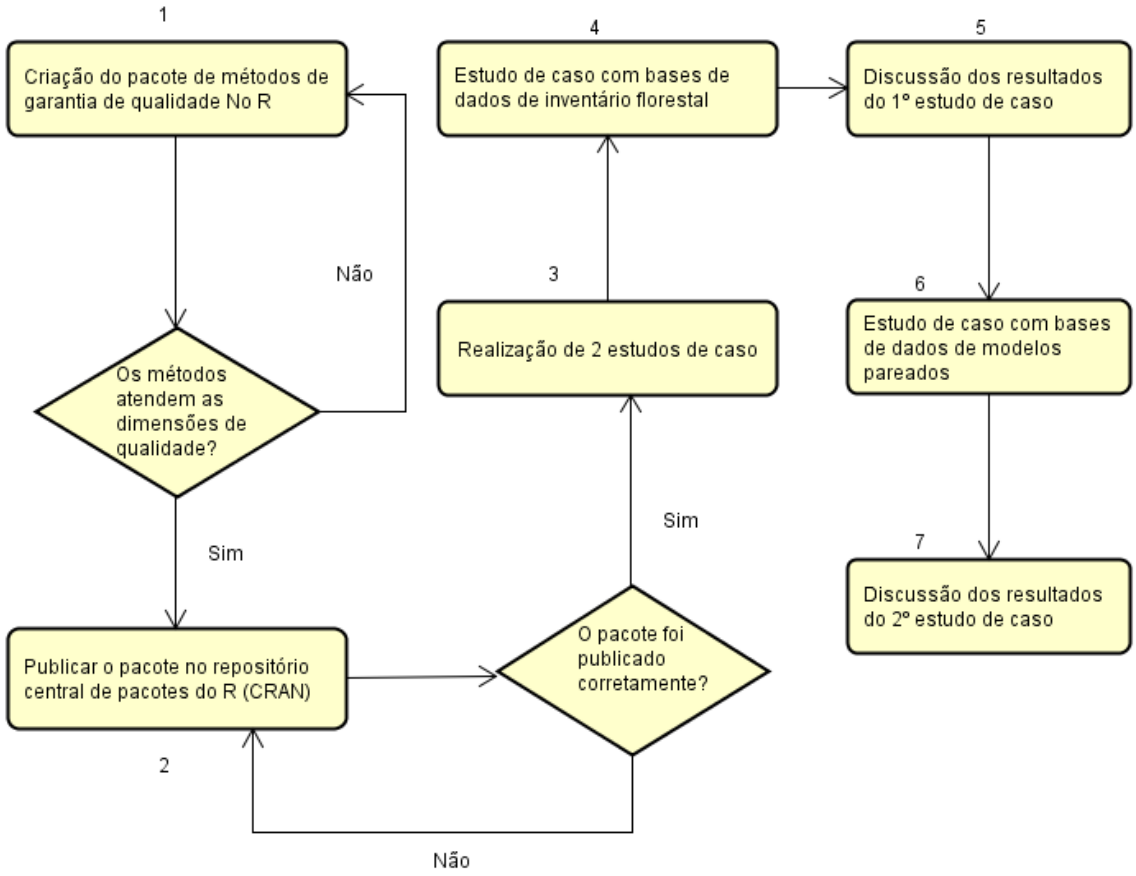


Figura 19 Fluxo de execução do trabalho.

A metodologia do trabalho se dá início em 1 criando o pacote de métodos de garantia de qualidade no R, após a implementação de todas as funções iremos verificar se os métodos em si criados atendem as dimensões de qualidade propostas no trabalho que são: consistência, pontualidade, completude. Uma vez validada essa etapa, ou seja, os erros encontrados durante os testes se encaixam nas dimensões propostas será realizada a publicação do pacote no repositório CRAN como apresentado em 2 na Figura 19, se o pacote for publicado corretamente então passamos para as próximas etapas (3,4,5,6,7).

As próximas etapas consistem na realização dos estudos de caso com bases de inventário florestal e bases de inventário florestal pareadas, neste momento é realizada a utilização do pacote afim de comprovar sua eficácia na identificação de erros nas bases. Contudo, após a realização de cada estudo de caso será realizada

uma discussão sobre os resultados, elaborando métricas para demonstrar o quão os erros encontrados afetam na base, e a quantidade de erros em relação ao total de registros na base utilizada.

4.1 Desenvolvimento do pacote para o R

Foi desenvolvido um pacote de métodos de garantia de qualidade de dados, o mesmo tem como objetivo implementar funções para detecção de não-conformidades em uma base de dados florestal.

Tais funções representam a implementação de procedimentos de consistência apresentados na seção 3.8 deste trabalho. Vale ressaltar que a qualidade de dados é alcançada no momento em que as funções criadas priorizam as dimensões de qualidade, dimensões essas que foram propostas pelos métodos e frameworks de gestão de dados apresentados na seção 3.2 deste trabalho. Em destaque a pontualidade, consistência, completude que são as principais dimensões avaliadas neste trabalho. As análises realizadas por meio dessas funções buscam possíveis erros em torno das dimensões relacionadas a qualidade dos dados, tendo em vista a função que verifica se uma árvore tem o DAP2 < DAP1, caso não existam os dados necessários para efetuar a comparação (DAP1 ou DAP2), conclui-se que a qualidade dos dados foi afetada, ou seja, alguma das dimensões não foi atendida, nesse caso especificamente a pontualidade, ou seja, os dados não estavam disponíveis quando foi necessário. Dessa forma, alia-se os procedimentos de consistência com as dimensões de qualidade a fim de atingir os objetivos propostos neste trabalho.

O pacote foi denominado *Forest Data Quality* (FDQ), o mesmo foi desenvolvido utilizando a linguagem R e fará parte do software estatístico R (R DEVELOPMENT CORE TEAM, 2005) que auxilia na manipulação de cálculos matemáticos e estatísticos. A Figura 20 apresenta o diagrama de atividades da criação das funções que irão compor o pacote FDQ.

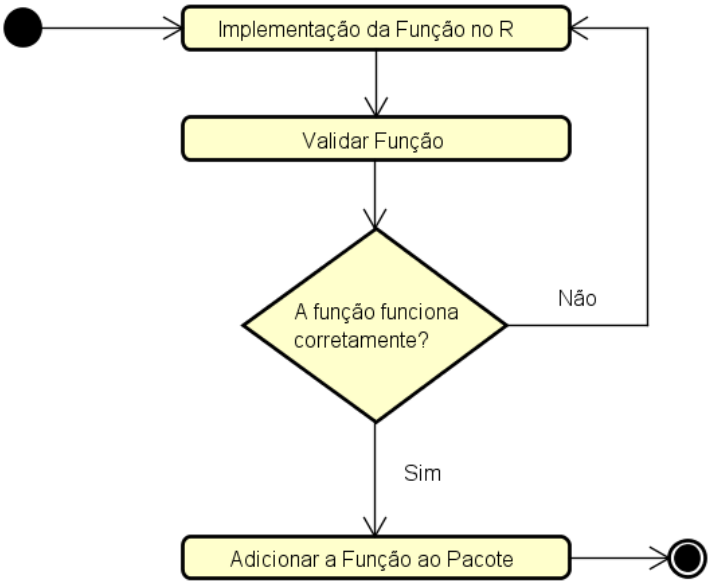


Figura 20 Implementação das funções no R.

A Figura 20 apresenta o processo de criação de uma função do pacote, primeiramente é feita a implementação da função no R, após isso valida-se o que foi criado, verificando-se o funcionamento da função. E se tudo estiver correto então a mesma é adicionada ao pacote FQD, senão é retornada para etapa de implementação para realizar as correções necessárias. O processo é repetido até que todas as funções do pacote estejam implementadas e validadas. O quadro 9 exemplifica a descrição de uma das funções que farão parte do pacote FDQ.

Método 1: Avalia o crescimento do DAP em toda dimensão longitudinal	
Função	Avalia o crescimento do DAP ao longo das medições.
Descrição	Espera-se que o DAP cresça para árvores normais.
Entrada	Base de dados, campo DAP na idade 1 e campo DAP na idade 2.
Origem	Base de dados pareada, criada a partir da base de dados de inventário.
Saídas	Listagem de registros de árvores que apresentam DAP2<DAP1, dentre todas as parcelas avaliadas.
Destino	Avaliação do usuário quanto ao procedimento a ser tomado. Além disso, registro das não conformidades encontradas pela aplicação deste método.
Ação	Em toda a base de dados, considerando todas as parcelas disponíveis, a função irá verificar a seguinte condição: Identificar árvores com DAP2 < DAP1 entre medições consecutivas i e i+1, sendo i o ano da medição.

Quadro 9 Descrição da função de verificação de DAP.

Fonte: Adaptado de Fraga Filho (2016)

O quadro 9 apresenta um resumo sobre a função de verificação de DAP, onde se têm campos desde nome, descrição, entradas, passando pelo escopo (saídas) até a ação que a função irá realizar. A documentação do pacote apresentada no apêndice A deste trabalho apresenta uma descrição semelhante para todas as funções do pacote.

Uma vez finalizada a implementação do pacote o intuito é adicioná-lo ao repositório de pacotes central do R (R DEVELOPMENT CORE TEAM, 2005) que é chamado de CRAN, dessa forma podemos disponibilizar o FDQ para qualquer usuário que deseja utilizar as funções presentes nele, a Figura 21 apresenta os passos necessários para submissão do pacote ao mesmo.

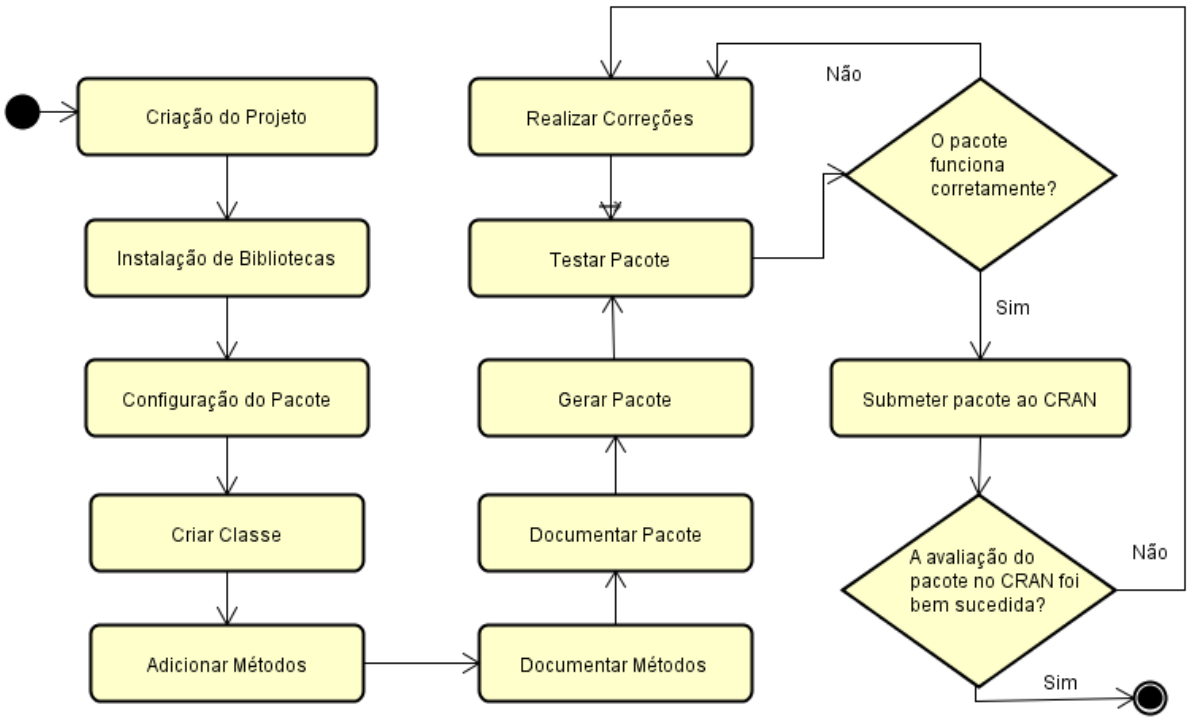


Figura 21 Diagrama de atividades para submissão do FDQ ao CRAN.

Inicialmente, é criado um projeto de submissão do pacote, após isso, se necessário, é realizada a importação das bibliotecas necessárias para o funcionamento do seu pacote. Depois são realizadas configurações acerca do pacote como descrição e informações pertinente ao pacote, para enfim ser criada uma classe onde serão adicionados todos os seus métodos (funções).

Feito isso é realizado toda documentação quanto aos métodos criados e ao pacote em geral. É realizada então a geração do pacote e os testes a fim de se validar o que foi criado, se o pacote funciona corretamente então é submetido ao CRAN, o mesmo irá realizar uma avaliação do pacote levando em conta características como código, idioma, parâmetros, documentação entre outros. Se a avalição for aprovada então o pacote é adicionado ao repositório deles e qualquer usuário do R poderá instalar o mesmo.

4.2 Descrição dos dados

Neste estudo foram utilizados dados de povoamentos não desbastados com 20 genótipos (clones de eucalipto) localizados na região Centro-oeste do estado de MG. O primeiro conjunto de dados apresenta plantio no espaçamento de 3 x 2 m, 3 x 2,5

m, 3 x 2,7 m, 3 x 2,8 m, 3 x 3 m, 3,3 x 2 m, 3,3 x 3 m, e 4 x 2,5 m. Os dados foram provenientes de 1752 parcelas permanentes, variando entre 274 e 356 m2, medidas entre 2006 e 2012. Nas análises foram utilizadas 1377 parcelas, 392 foram medidas aos 2 anos, 349 em 3 anos, 339 em 4 anos, 240 em 5 anos, 39 em 6 anos, e 18 em 7 anos.

No segundo conjunto de dados o plantio se deu no espaçamento 3 x 3 m. Os dados foram provenientes de 376 parcelas permanentes, variando entre 274 e 356 m2, medidas entre 2006 e 2012. Nas análises foram utilizadas 376 parcelas das quais foram medidas em 4 anos, 155 parcelas medidas em 5 anos, 23 parcelas medidas em 6 anos e 13 parcelas com 7 medições, todas consecutivas.

4.3 Processamento dos dados

Todo processamento dos dados, criação do pacote e geração de resultados foram realizados na linguagem R (R DEVELOPMENT CORE TEAM, 2005). A escolha se deu pela possibilidade de manipulação de grandes volumes de dados e a eficácia no processamento, além de todo apoio de funções estatísticas que o R possui.

A manipulação dos dados e geração de resultados, gráficos foi realizada utilizando os pacotes da linguagem R: *sqldf* (GROTHENDIECK, 2014), *data.table* (DOWLE et al., 2014), *plyr* (WICKHAM, 2011), *Fgmutils* (FRAGA FILHO et al., 2015), *ggplot2* (WICKHAM, 2016), *randomcoloR* (AMMAR, 2016).

5 RESULTADOS E DISCUSSÃO

Tendo em vista a metodologia apresentada foi feito o desenvolvimento do pacote *Forest Data Quality* no R baseando-se nos procedimentos de detecção de inconsistência apresentados na revisão de literatura deste trabalho propostos por Fraga Filho (2016) e Demolinari (2006). Após o desenvolvimento do pacote realizou-se dois estudos de caso afim de verificar a eficácia do pacote criado na identificação de erros relacionados a qualidade de dados.

Para cada estudo de caso foram realizados testes com as nove funções de análise contidas no pacote para bases com idades entre 24 e 96 meses. A escolha da idade é justificada pela presença dos dados disponíveis, os mesmos são frequentemente utilizados nas empresas florestais no ajuste de modelos para obtenção de resultados de estimativas para o setor. Além disso, a primeira medição é realizada de 24 meses e aos 96 meses é o que se considera para efeito de colheita, entre 6 e 8 anos.

5.1 Primeiro Estudo de Caso

A primeira base de dados analisada foi de dados de inventário florestal, a mesma continha 179343 registros, com 29 variáveis (colunas), 1752 parcelas, 3 sítios. As análises descritas em 5.1.1,5.1.2, 5.1.3 exigem que a base esteja pareada, uma base pareada é muito utilizada em modelos de crescimento e produção, pois, pode-se ajustar modelos com base em medições consecutivas de um ano para o outro, para variáveis de mensuração como é o caso das três primeiras análises que iremos realizar. As Figuras 22 e 23 representam uma base não pareada e uma pareada respectivamente.

	cod_id	ano_medicao	dap	ht	estado
1	10_1394_10_71__2_1	2011	9.9	14.44	N
2	10_1394_10_71__2_1	2012	11.4	16.87	N
3	10_1394_11_71__2_1	2011	12.7	16.76	N
4	10_1394_11_71__2_1	2012	15.7	19.77	N
5	10_1394_12_71__2_1	2011	9.9	14.44	N
6	10_1394_12_71__2_1	2012	12.4	17.59	N
7	10_1394_12_71__2_2	2011	9.9	14.44	N
8	10_1394_12_71__2_2	2012	12.2	17.45	N
9	10_1394_13_71__3_1	2011	12.1	16.29	N
10	10_1394_13_71__3_1	2012	15.0	19.33	N

Figura 22 Exemplo de base não pareada.

	cod_id	ano_medicao1	ano_medicao2	dap1	dap2	ht1	ht2	estado
1	10_1394_10_71__2_1	2011	2012	9.9	11.4	14.44	16.87	N
2	10_1394_11_71__2_1	2011	2012	12.7	15.7	16.76	19.77	N
3	10_1394_12_71__2_1	2011	2012	9.9	12.4	14.44	17.59	N
4	10_1394_12_71__2_2	2011	2012	9.9	12.2	14.44	17.45	N
5	10_1394_13_71__3_1	2011	2012	12.1	15.0	16.29	19.33	N
6	10_1394_14_71__3_1	2011	2012	11.1	14.6	15.47	19.08	N
7	10_1394_15_71__3_1	2011	2012	10.7	13.8	15.13	18.55	N
8	10_1394_16_71__3_1	2011	2012	11.4	13.7	15.72	18.48	N
9	10_1394_17_71__3_1	2011	2012	12.7	15.7	16.76	19.77	N
10	10_1394_18_71__3_1	2011	2012	12.5	15.4	16.60	19.58	N

Figura 23 Exemplo de base de dados pareada.

Na Figura 22 temos o exemplo de uma base de dados que não contém campos pareados, já na Figura 23 temos a mesma base, porém com o campo “dap” pareado que se tornou “dap1” e “dap2” que tem respectivos anos de medição (ano_medicao1, ano_medicao2), percebe-se outras variáveis que foram pareadas durante o processo como “ht”, vale ressaltar que neste caso o campo de comparação foi “ano_medicao”, portanto não deve ser considerada uma variável pareada.

Contudo para obter a base pareada foi utilizada a função criarDadosPareados do pacote *Fgmutils* (FRAGA FILHO et al., 2015) disponibilizado no repositório da linguagem R que é o CRAN. Após o pareamento obteve-se 110377 registros e 1375 parcelas, uma vez obtida a base pareada, a mesma foi submetida aos métodos do pacote e o resultado é apresentado abaixo.

5.1.1 Método de análise: *check_measurement_ages*

Esta análise verifica se uma determinada variável de medição possui registros do tipo variável na idade 2 menores que variável na idade 1 em idades consecutivas i+1 e i, nesta função foram avaliadas as variáveis de mensuração “DAP” e “HT”, a Figura 24 apresenta exemplos dos resultados obtidos para primeira variável testada.

	cod_id	parcela	dap1	ano_medicao1	dap2	ano_medicao2	estado
1	10_1397_2_72__1_1	1397	11.4	2011	0.0	2012	M
2	10_1398_12_73__2_2	1398	6.7	2011	6.6	2012	N
3	10_1398_14_73__3_2	1398	7.6	2011	7.5	2012	N
4	10_155_14_84__2_1	155	5.9	2006	5.8	2007	I
5	10_155_17_84__3_1	155	17.1	2006	16.8	2007	N
6	10_155_18_84__3_1	155	13.2	2006	13.0	2007	N
7	10_155_1_84__1_1	155	12.9	2006	12.6	2007	N
8	10_155_21_84__3_1	155	16.9	2006	16.8	2007	N
9	10_155_35_84__5_1	155	12.2	2006	12.1	2007	N
10	10_155_46_84__7_1	155	19.4	2006	19.3	2007	N

Figura 24 Resultados da avaliação do DAP.

A Figura 24 nos permite visualizar os 10 primeiros registros da encontrados errados na base, e podemos ver medições da variável “dap2” menores que a “dap1”. A análise resultou em 3929 registros com erros, correspondendo a 3,57% do total de registros da base, além disso 62,32% das parcelas apresentam problemas desse tipo. A Figura 25 mostra a mesma análise realizada, porém para variável de medição “ht”.

	cod_id	parcela	ht1	ano_medicao1	ht2	ano_medicao2	estado
1	10_1397_2_72__1_1	1397	15.70	2011	0.00	2012	M
2	10_155_13_84__2_1	155	27.10	2006	25.50	2007	N
3	10_155_14_84__2_1	155	13.90	2006	0.00	2007	I
4	10_155_19_84__3_1	155	25.90	2006	25.60	2007	N
5	10_155_1_84__1_1	155	25.10	2006	0.00	2007	N
6	10_155_29_84__4_1	155	28.90	2006	28.70	2007	D
7	10_155_40_84__6_1	155	25.50	2006	24.90	2007	N
8	10_156_13_84__2_1	156	29.30	2006	0.00	2007	N
9	10_156_4_84__1_1	156	21.10	2006	0.00	2007	N
10	10_157_22_84__4_1	157	20.50	2006	0.00	2007	N

Figura 25 Resultados da avaliação HT.

A Figura 25 contém exemplos de registros com erros para variável “ht1” e “ht2”, foram encontrados 3918 registros com erros que correspondem a 3,54% do total de registros com erros na base, a porcentagem de parcelas com registros com erros é 62,25%

5.1.1 Método de análise: *check_ages*

Esta análise verifica se existem idades classe de idade em idades consecutivas com diferença maior que 12 para verificação utilizando mês, e maior que um para verificação utilizando ano, o parâmetro utilizado para teste foi mês e a Figura 26 apresenta exemplos de resultados encontrados.

	cod_id	parcela	idadearred1	ano_medicao1	idadearred2	ano_medicao2
1	10_174_42_79__7_0	174	12	2006	36	2008
2	12_904_10_91__2_0	904	12	2008	36	2011
3	12_904_11_91__2_0	904	12	2008	36	2011
4	12_904_23_91__3_0	904	12	2008	36	2011
5	12_904_26_91__4_0	904	12	2008	36	2011
6	12_904_27_91__4_0	904	12	2008	36	2011
7	12_904_28_91__4_0	904	12	2008	36	2011
8	12_904_2_91__1_0	904	12	2008	36	2011
9	12_904_31_91__4_0	904	12	2008	36	2011
10	12_904_32_91__4_0	904	12	2008	36	2011

Figura 26 Resultado analise *check_ages*.

A Figura 26 contém exemplos de registros com erros para variável “idadearred1” e “idadearred2”, foram encontrados 765 registros com erros que correspondem a 0,69% do total de registros com erros na base, a porcentagem de parcelas com registros com erros é 6,25%.

5.1.2 Método de análise: *check_parcel_different_spacing*

Esta análise verifica se existem parcelas com diferentes espaçamentos em idades subsequentes i e i+1, a base pareada foi submetida a esta análise, porém não foram encontrados erros.

Por fim, finalizadas as análises que necessitavam da base pareada para serem utilizadas voltamos a utilizar nossa base original para realização da análise de 5.1.4 até 5.1.9.

5.1.3 Método de análise: *check_clone_different_plot*

Esta análise verifica se existem clones com diferentes parcelas, a base foi submetida a análise, porém não foram encontrados erros.

5.1.4 Método de análise: *check_dead_state*

Esta análise verifica se existem medições na (s) variável (s) informada com estado igual a M, no teste realizado as variáveis submetidas foram DAP, HT, a Figura 27 apresenta exemplos dos resultados encontrados.

	cod_id	parcela	estado	dap	ht
46	9_362_43_70__4_1	362	M	6.20	0.00
47	16_743_23_138__4_2	743	M	6.20	15.41
48	85_46_7_603__1_2	46	M	6.30	0.00
49	90_202_46_664__7_2	202	M	6.30	0.00
50	73_826_27_537__5_2	826	M	6.30	15.67
51	85_47_25_604__4_1	47	M	6.40	0.00
52	90_198_43_662__6_2	198	M	6.40	0.00
53	90_208_23_665__4_2	208	M	6.40	0.00
54	85_47_9_604__2_2	47	M	6.50	0.00
55	78_73_29_561__5_1	73	M	6.50	0.00
56	90_211_39_667__6_2	211	M	6.50	0.00

Figura 27 Resultados da análise *check_dead_state*.

A Figura 27 contém exemplos de registros com erros, as variáveis de medição apresentaram medições mesmo estando com estado igual a MORTO (M). Foram encontrados 267 registros com erros que correspondem a 0,14% do total de registros com erros na base, a porcentagem de parcelas com registros com erros é 5,07%.

5.1.5 Método de análise: *check_measurement_state*

Esta análise verifica se a (s) variável (s) de mensuração possuem medições iguais a zero nos estados diferentes de M (morta), F (falha), A (dominada). As variáveis impostas ao teste foram DAP e HT, a Figura 28 apresenta exemplos de resultados encontrados.

	cod_id	parcela	estado	dap	ht
1	85_45_48_602__6_2	45	N	0.0	0
2	85_54_43_607__6_1	54	N	0.0	0
3	10_176_3_80__1_2	176	N	0.0	0
4	47_415_7_360__1_1	415	N	0.0	0
5	47_416_55_360__7_1	416	N	0.0	0
6	47_416_9_360__2_1	416	N	0.0	0
7	45_549_38_331__5_2	549	I	0.0	0
8	88_566_24_634__3_1	566	I	0.0	0
9	27_585_9_215__2_2	585	N	0.0	0
10	46_793_23_335__5_2	793	C	0.0	0

Figura 28 Resultados da análise *check_measurement_state*.

A Figura 28 contém exemplos de registros com erros, as variáveis de medição (dap,ht) apresentaram medições iguais a 0 porém seus estados foram diferentes de M,F,A. Foram encontrados 72886 registros com erros que correspondem a 40,64% do total de registros com erros na base, a porcentagem de parcelas com registros com erros é 75,11%.

5.1.6 Método de análise: *check_undefined_spacing*

Está análise verifica se existem espaçamentos que não foram definidos dentro da base de dados, se são iguais a 0 ou possuem valores nulos, a Figura 29 apresenta exemplos de resultados encontrados.

	cod_id	parcela	espacamento
1	79_989_3_568__1_1	989	NA
2	79_989_12_568__2_1	989	NA
3	79_989_11_568__2_1	989	NA
4	79_989_9_568__2_1	989	NA
5	79_989_20_568__3_1	989	NA
6	79_989_28_568__4_1	989	NA
7	79_989_19_568__3_1	989	NA
8	79_989_4_568__1_1	989	NA
9	79_989_2_568__1_1	989	NA
10	79_989_8_568__2_1	989	NA

Figura 29 Resultados análise *check_undefined_spacing.*

A Figura 29 contém exemplos de registros com erros, o campo espaçamento está com valores indefinidos (nulos). Foram encontrados 112 registros com erros que correspondem a 0,062% do total de registros com erros na base, a porcentagem de parcelas com registros com erros é 0,00019%.

5.1.7 Método de análise: *check_zero_measurement*

Esta análise verifica quais variáveis de medição possuem valores igual a 0 nos estados que o usuário informou, as variáveis utilizadas foram DAP, HT, e o estado a ser verificado foi N (árvore normal). A Figura 30 apresenta exemplos de resultados encontrados.

	cod_id	parcela	estado	dap	ht
1	85_45_48_602__6_2	45	N	0.0	0
2	85_54_43_607__6_1	54	N	0.0	0
3	10_176_3_80__1_2	176	N	0.0	0
4	47_415_7_360__1_1	415	N	0.0	0
5	47_416_55_360__7_1	416	N	0.0	0
6	47_416_9_360__2_1	416	N	0.0	0
7	27_585_9_215__2_2	585	N	0.0	0
8	98_1105_19_702__3_2	1105	N	0.0	0
9	98_1108_13_703__2_1	1108	N	0.0	0
10	98_1108_20_703__3_1	1108	N	0.0	0

Figura 30 Resultado análise *check_zero_measurement*.

A Figura 30 contém exemplos de registros com erros o campo estado consta como N e existem variáveis de medição como “dap” e “ht” com valores iguais a 0. Foram encontrados 66072 registros com erros que correspondem a 36,84% do total de registros incorretos na base, a porcentagem de parcelas com registros com erros com medições iguais a zero e estado igual a N é de 75%.

5.1.8 Interseção de inconsistências

Tendo em vista a aplicação dos oito primeiros métodos de análise, foi realizada a interseção entre cada resultado encontrado por cada método afim para saber quais registros possuíam mais de algum tipo de inconsistência.

***check_measurement_ages (“dap”, “ht”)*: 7847 registros**	
Método de Análise	**Quantidade de registros encontrados**
check_ages	84
check_dead_state	141
check_measurement_state	3609
check_undefined_spacing	1
check_zero_measurement	2611

Quadro 10 Interseção dos resultados da função *check_measurement_ages*

O quadro 10 apresenta os resultados obtidos (7847) na função *check_measurement_ages* aplicados aos demais métodos de análise que obtiveram algum resultado, a coluna quantidade de registros apresenta o total de registro inconsistentes detectados pelas suas respectivas funções (método de análise), como observado acima todas as funções aplicadas ao resultado apresentaram algum (s) registros inconsistentes. Vale ressaltar que o script de como foi obtida as interseções dos resultados e um exemplo de base após interseção está disponível no apêndice A deste trabalho, em específico na seção 8.1.1.

Os quadros 11,12,13,14,15 apresentam a aplicação do resultado de cada método aos demais afim de encontrar interseções de inconsistências.

***check_ages*: 765 registros**	
Método de Análise	**Quantidade de registros encontrados**
check_measurement_ages ("dap"," ht")	70
check_dead_state	6
check_measurement_state	289
check_undefined_spacing	0
check_zero_measurement	267

Quadro 11 Interseção dos resultados da função *check_ages.*

***check_dead_state*: 267 registros**	
Método de Análise	**Quantidade de registros encontrados**
check_measurement_ages ("dap"," ht")	115
check_ages	6
check_measurement_state	0
check_undefined_spacing	0
check_zero_measurement	0

Quadro 12 Interseção dos resultados da função *check_ages.*

***check_measurement_state*: 72886 registros**	
Método de Análise	**Quantidade de registros encontrados**
check_measurement_ages ("dap"," ht")	3532
check_ages	289
check_dead_state	0
check_undefined_spacing	54
check_zero_measurement	66072

Quadro 13 Interseção dos resultados da função *check_measurement_state.*

***check_undefined_spacing*: 112 registros**	
Método de Análise	**Quantidade de registros encontrados**
check_measurement_ages ("dap"," ht")	1
check_ages	0
check_dead_state	0
check_measurement_state	54
check_zero_measurement	52

Quadro 14 Interseção dos resultados da função *check_undefined_spacing.*

***check_zero_measurement*: 112 registros**	
Método de Análise	**Quantidade de registros encontrados**
check_measurement_ages ("dap"," ht")	2558
check_ages	267
check_dead_state	0
check_measurement_state	66072
check_undefined_spacing	52

Quadro 15 Interseção dos resultados da função *check_measurement_state.*

Por meio dos quadros 11, 12, 13, 14, 15 é possível identificar inconsistências além das detectadas pelo método em si, contribuindo desta forma para encontrar mais erros que afetam diretamente na qualidade dos dados, para realizar essa intersecção utilizamos de uma função nativa do R (R DEVELOPMENT CORE TEAM, 2005) chamada merge, no qual ele faz um chaveamento de dos dois resultados das duas funções pelo "cod_id" de cada árvore e retorna uma base como resultado.

5.1.9 Método de análise: *check_diametric_increase*

O método a seguir é uma adaptação do incremento diamétrico proposto por Demolinari (2006), o mesmo permite visualizar a dinâmica crescimento de árvores por parcela através de gráficos de barras coloridas onde cada cor representa uma classe de diâmetro, são gerados gráficos para cada idade e pode-se perceber pelo gráfico a mudança de classes de diâmetro das árvores.

No eixo x do gráfico são apresentadas as classes de diâmetro, caso a base não possua esse campo de classe o pacote *Forest Data Quality* permite a criação dessas classes por meio da função *generate_diameter_classes*, mais informações sobre ela

pode-se conferir a documentação do pacote apresentada no apêndice A deste trabalho. No eixo y do gráfico é apresentado o número árvores sobreviventes por hectare (N/Ha), na função você pode designar o campo que deseja aplicar para y porém para esta análise foi seguido o que foi proposto por Demolinari (2006).

Além disso, como escopo dessa função também é apresentado um quadro com a quantidade de saltos da parcela nas idades complementando ainda mais a análise. Vale salientar que essa análise não detecta diretamente como as outros, mas permite que o usuário visualize o crescimento e a partir disso detecte alguma anomalia perante os dados ali apresentados, por exemplo uma classe que cresceu demais ou não cresceu durante um determinado período de tempo pode significar que existe dados errados naquela parcela.

No estudo de caso em questão foi analisada a parcela com mais registros dentro da base com intuito de poder visualizar as mudanças dentre as classes, os resultados são apresentados nas Figuras 31 a 35.

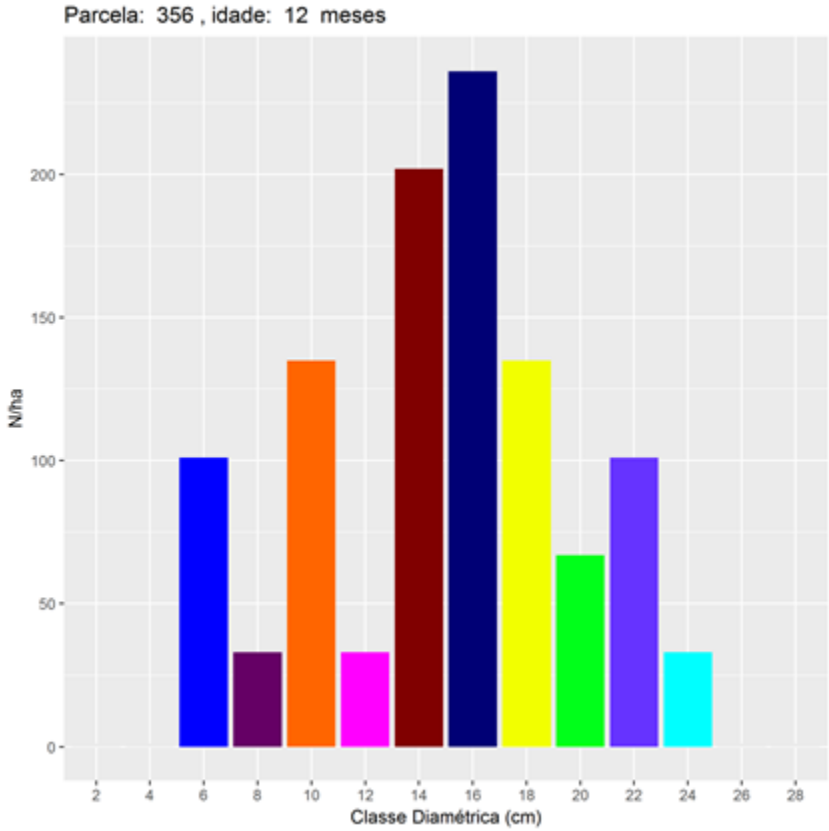


Figura 31 Incremento Diamétrico da parcela 356 aos 12 meses.

Na Figura 31 é apresentada a parcela 356 pertencente ao sítio 21 aos 12 meses e sua classe de produtividade baixa. Quanto a classe de produtividade, Demolinari (2006) afirma que de uma maneira geral, observou-se, para todas as idades, que, independente da classe de produtividade, quanto maior a classes diamétrica, maior é a taxa de crescimento, de tal forma que as maiores árvores, em uma idade inicial, atingiram classes de tamanho superiores em idades futuras (SOUZA et al., 1993).

Com relação à capacidade produtiva, constatou-se haver uma tendência inversa com momento de estagnação do crescimento, principalmente das menores árvores, ou seja, quanto melhor o local mais cedo se verifica esta estagnação do crescimento. Dessa forma como o sítio 21 pertence a uma classe de produtividade baixa a tendência das árvores da parcela 356 a estagnação do crescimento não é rápida, portanto espera-se ver muitas mudanças de classe durante as idades.

Nesta parcela temos classes de diâmetro variando de 6 a 24 cm de diâmetro, como 12 é a primeira medição não há variações de classes de diâmetro ainda.

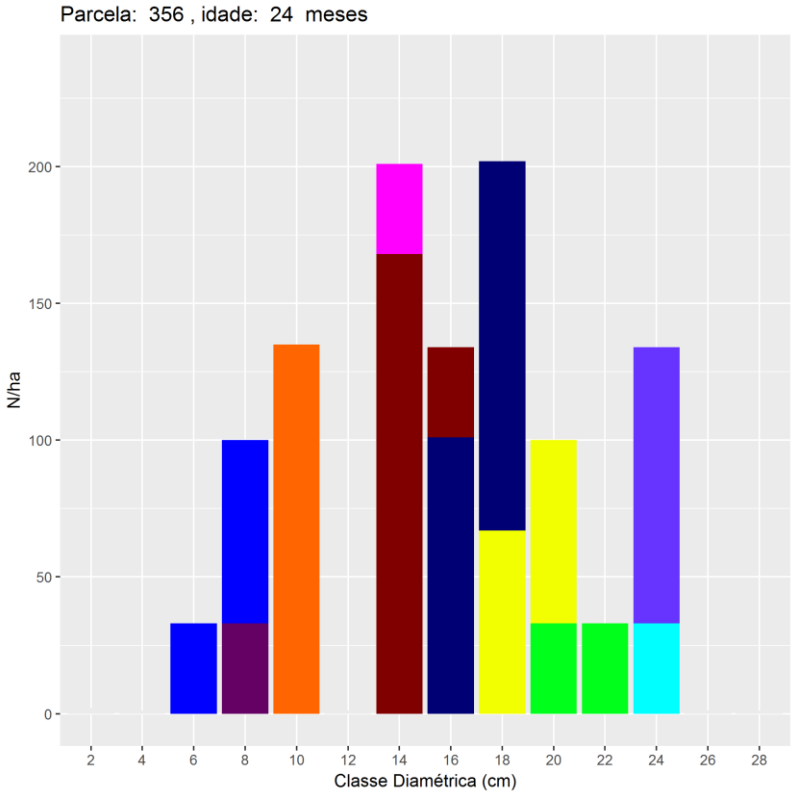


Figura 32 Incremento Diamétrico da parcela 356 aos 24 meses.

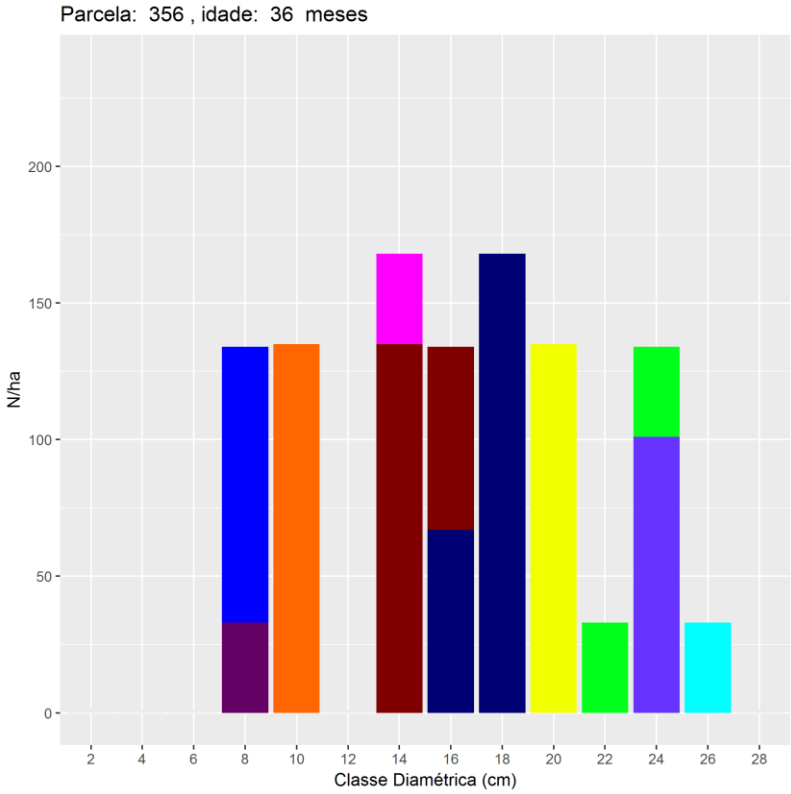


Figura 33 Incremento Diamétrico da parcela 356 aos 36 meses.

Pode-se perceber que da idade 24 para 36 que o restante das árvores que estavam na classe 6 foram para classe 8, e o restante da classe 18 foi para 20. Porém a classe 10 continuou estagnada desde a idade inicial, podendo assim, ser um objeto de estudo para uma análise mais aprofundada afim de detectar alguma inconformidade sobre os dados.

As Figuras 32 e 33 apresentam as medições da parcela 356 para 24 e 36 meses respectivamente, através das cores de cada classe de diâmetro pode-se perceber as mudanças durante as diferentes medições. As Figuras 34 e 35 apresentam o processo para as duas últimas medições realizadas na parcela 356.

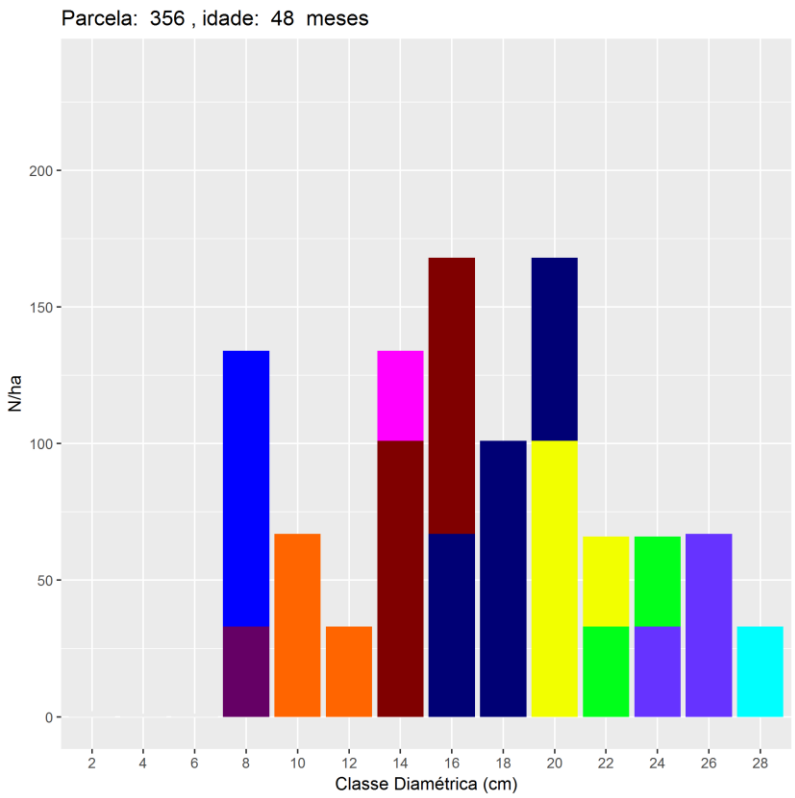


Figura 34 Incremento Diamétrico da parcela 356 aos 48 meses.

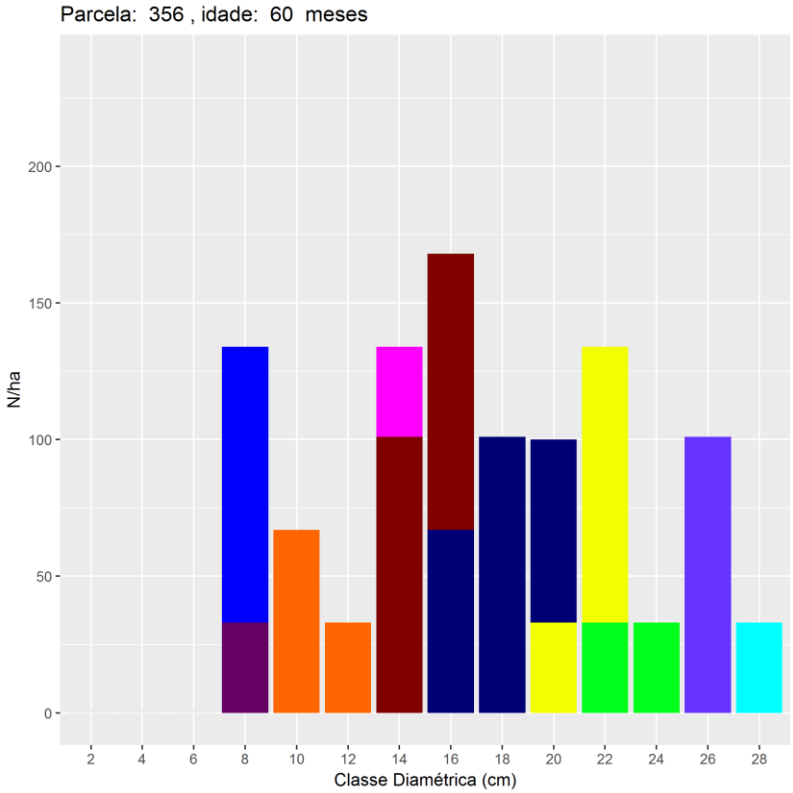


Figura 35 Incremento Diamétrico da parcela 356 aos 60 meses.

A Figura 36 apresentada abaixo contém 10 registros da tabela de saltos da parcela 356 por idade, a mesma também é um escopo da função *check_diametric_increase*, permitindo assim o usuário ter mais uma forma de analisar o crescimento da parcela. É importante ressaltar que a tabela apresenta o salto por árvore ao invés do salto da parcela toda, permitindo assim uma visão detalhada da dinâmica de crescimento dentro da parcela. Vale salientar que cada árvore dentro de uma base têm um identificador único para análises unitárias em modelos de árvore individual, na base em questão este identificador é representando na Figura abaixo pela coluna *cod_id.*

	cod_id	age_24	age_36	age_48	age_60
1	9_356_0_67__0_0	2	0	2	0
2	9_356_10_67__1_0	0	0	2	0
3	9_356_10_67__1_1	0	0	2	0
4	9_356_11_67__1_1	0	0	0	0
5	9_356_12_67__1_0	0	0	2	0
6	9_356_12_67__1_1	0	0	2	0
7	9_356_13_67__2_0	0	0	2	0
8	9_356_13_67__2_1	0	0	2	0
9	9_356_14_67__2_1	2	0	0	0
10	9_356_15_67__2_1	2	0	2	0

Figura 36 Saltos de classes parcela 356.

Todavia, o número 0 significa que não houve salto naquela idade, e os números restantes representam a quantidade de saltos de classe de diâmetro ocorridos naquela idade.

Vale ressaltar que a imagem 36 mostra os saltos a partir da idade 24 e não de 12 pois é a primeira idade que se pode analisar saltos, como 12 é a primeira medição não faz sentido ter saltos de uma idade anterior (inexistente) para ela. Tomando como exemplo a primeira linha da Figura 36 pode-se visualizar que na idade 24 a árvore teve um salto de 2 classes, na idade 36 não houve saltos, e assim por diante. Contudo, a verificação desta tabela pode servir para identificar inconformidades de dados, analisando o salto de uma árvore e questionando se ela deveria realizar esse salto de classe de uma idade para outra, encaixando-se assim como uma forma de identificar problemas relacionados a qualidade dos dados.

Por fim, mediante a realização do incremento diamétrico pode se acompanhar completamente o desenvolvimento das árvores na parcela desde a medição inicial até a medição final, os gráficos e a tabela contendo os saltos servem de base para identificação de possíveis erros e tomadas de decisões que possam surgir em relação a parcela, o profissional que utilizará a função poderá realizar tomadas de decisões baseadas nos resultas ou aferir por erros mais internos utilizando as funções do FDQ.

5.2 Segundo Estudo de Caso

O segundo estudo de caso também foi direcionado a uma base de dados de inventário florestal, a mesma foi submetida a todos métodos que foram aplicados na primeira base.

A segunda base de dados analisada possui 52006 registros, com 39 variáveis (colunas), 376 parcelas, 3 sítios. As análises descritas em 5.2.1,5.2.2, 5.2.3 exigem que a base esteja pareada, para isso foi realizado o mesmo procedimento de pareamento da base do estudo de caso anterior. Quanto as funções de 5.2.4 a 5.2.9 não necessitam de a base estar pareada, os resultados são apresentados abaixo.

5.2.1 Método de análise: *check_measurement_ages*

Nesta função foram avaliadas as variáveis de mensuração “DAP” e “HT”, a Figura 37 apresenta exemplos dos resultados obtidos para primeira variável testada.

	cod_id	parcela	dap1	ano_medicao1	dap2	ano_medicao2	estado
1	10_23_21_42__4_1	23	17.1	2008	16.8	2009	A
2	10_23_2_42__1_1	23	13.7	2008	13.5	2009	A
3	10_23_30_42__5_1	23	13.3	2006	12.8	2007	A
4	10_24_25_42__5_1	24	11.2	2006	0.0	2007	Q
5	10_25_11_43__2_1	25	15.0	2006	0.0	2007	Q
6	10_25_12_43__2_1	25	16.2	2006	0.0	2007	Q
7	10_25_14_43__3_1	25	13.8	2006	0.0	2007	Q
8	10_25_18_43__3_1	25	17.1	2006	0.0	2007	Q
9	10_25_23_43__4_1	25	13.6	2006	0.0	2007	Q
10	10_25_25_43__5_1	25	16.2	2006	0.0	2007	Q

Figura 37 Resultado da análise *check_measurement_ages* para variável DAP.

A Figura 37 nos permite visualizar os 10 primeiros registros da encontrados errados na base, e podemos ver medições da variável “dap2” menores que a “dap1”. A análise resultou em 654 registros com erros, correspondendo a 1,62% do total de registros da base, além disso 68,88% das parcelas apresentam problemas desse tipo.

A Figura 38 mostra a mesma análise realizada, porém para variável de medição “ht”.

	cod_id	parcela	ht1	ano_medicao1	ht2	ano_medicao2	estado
1	10_24_25_42__5_1	24	16.71117	2006	0.00000	2007	Q
2	10_25_11_43__2_1	25	19.03923	2006	0.00000	2007	Q
3	10_25_12_43__2_1	25	19.71527	2006	0.00000	2007	Q
4	10_25_14_43__3_1	25	18.28386	2006	0.00000	2007	Q
5	10_25_18_43__3_1	25	20.17025	2006	0.00000	2007	Q
6	10_25_23_43__4_1	25	18.15025	2006	0.00000	2007	Q
7	10_25_25_43__5_1	25	19.71527	2006	0.00000	2007	Q
8	10_25_7_43__2_1	25	19.33056	2006	0.00000	2007	Q
9	10_26_22_43__4_1	26	17.00000	2006	0.00000	2007	Q
10	10_26_5_43__1_1	26	15.00000	2006	0.00000	2007	Q

Figura 38 Resultado da análise *check_measurement_ages* para variável HT.

A Figura 38 nos permite visualizar os 10 primeiros registros da encontrados errados na base, e podemos ver medições da variável “ht2” menores que a “ht1”. A análise resultou em 779 registros com erros, correspondendo a 1,93 % do total de registros da base, além disso 57,44% das parcelas apresentam problemas desse tipo.

5.2.2 Método de análise: *check_ages*

	cod_id	parcela	idadearred1	ano_medicao1	idadearred2	ano_medicao2
1	1_224_10_4__2_1	224	24	2007	48	2009
2	1_224_11_4__2_1	224	24	2007	48	2009
3	1_224_12_4__2_1	224	24	2007	48	2009
4	1_224_13_4__3_1	224	24	2007	48	2009
5	1_224_14_4__3_1	224	24	2007	48	2009
6	1_224_15_4__3_1	224	24	2007	48	2009
7	1_224_16_4__3_1	224	24	2007	48	2009
8	1_224_17_4__3_1	224	24	2007	48	2009
9	1_224_18_4__3_1	224	24	2007	48	2009
10	1_224_19_4__4_1	224	24	2007	48	2009

Figura 39 Resultado da análise *check_ages*.

A Figura 39 contém exemplos de registros com erros para variável “idadearred1” e “idadearred2”, foram encontrados 322 registros com erros que correspondem a 0,79% do total de registros com erros na base, a porcentagem de parcelas com registros com erros é 10,63%.

5.2.3 Método de análise: *check_parcel_different_spacing*

Esta análise verifica se existem parcelas com diferentes espaçamentos em idades subsequentes i e i+1, a base pareada foi submetida a esta análise, porém não foram encontrados erros.

5.2.4 Método de análise: *check_clone_different_plot*

Esta análise verifica se existem clones com diferentes parcelas, a base foi submetida a análise, porém não foram encontrados erros.

5.2.5 Método de análise: *check_dead_state*

Esta análise verifica se existem medições na (s) variável (s) informada com estado igual a M, para isto foram realizadas análises nas variáveis de mensuração DAP, HT. Porém, não foram encontrados erros relacionados a esta análise na base de dados.

5.2.6 Método de análise: check_measurement_state

As variáveis que foram submetidas a avalição nesta análise foram DAP, HT, os resultados são apresentados na imagem abaixo.

	cod_id	parcela	estado	dap	ht
1	2_1_7_9__2_1	1	Q	0	0
2	2_1_2_9__1_1	1	Q	0	0
3	2_1_7_9__2_1	1	Q	0	0
4	13_6_1_68__1_1	6	Q	0	0
5	31_16_19_143__3_1	16	Q	0	0
6	31_16_19_143__3_1	16	Q	0	0
7	31_16_19_143__3_1	16	Q	0	0
8	31_16_19_143__3_1	16	Q	0	0
9	31_21_31_144__5_1	21	Q	0	0
10	31_22_14_142__2_1	22	Q	0	0

Figura 40 Resultado da análise *check_measurement_state.*

A Figura 40 contém exemplos de registros com erros, as variáveis de medição (“dap”, ”ht”) apresentaram medições iguais a zero, porém, seus estados foram diferentes de M, F, A. Foram encontrados 1040 registros com erros que correspondem a 1,99% do total de registros com erros na base, a porcentagem de parcelas com registros com erros é 23,93%.

5.2.7 Método de análise: *check_undefined_spacing*

Está análise verifica se existem espaçamentos que não foram definidos dentro da base de dados, se são iguais a 0 ou possuem valores nulos, porém feita a análise não foram encontrados erros relacionados a análise.

5.2.8 Método de análise: *check_zero_measurement*

Esta análise verifica quais variáveis de medição possuem valores igual a 0 nos estados que o usuário informou, as variáveis utilizadas foram DAP, HT, e o estado a ser verificado foi N (árvore normal).

Para análise utilizando N como parâmetro de estado não foram encontrados registros, os estados encontrados com as variáveis de medição iguais a 0 foram “M”, “F”, “Q”, “A”, ou seja, não deveriam existir medições, portanto a análise não encontrou resultados.

5.2.9 Interseção de inconsistências

Assim como feito na seção 5.1.8 para o primeiro estudo de caso, após realizar as análises com a segunda base foram feitas as interseções dos resultados afim de encontrar possíveis inconsistências dentre os resultados já obtidos. Vale salientar que o segundo estudo de caso apenas 3 dos métodos obtiveram resultados, os quadros 16,17,18,19,20,21 apresentam os resultados das interseções.

***check_measurement_ages (“dap”, “ht”)*: 1433 registros**	
Método de Análise	**Quantidade de registros encontrados**
check_ages	44
check_measurement_state	240

Quadro 16 Interseção dos resultados da função *check_measurement_ages*

***check_ages*: 322 registros**	
Método de Análise	**Quantidade de registros encontrados**
check_measurement_ages (“dap”,” ht”)	22
check_measurement_state	3

Quadro 17 Interseção dos resultados da função *check_ages*

***check_measurement_state*: 1040 registros**	
Método de Análise	**Quantidade de registros encontrados**
check_measurement_ages (“dap”,” ht”)	240
check_ages	6

Quadro 18 Interseção dos resultados da função *check_measurement_state*

5.2.10 Método de análise: *check_diametric_increase*

Nesse estudo de caso foi analisada a parcela 149 do sítio 27 com classe de produtividade igual a média, a mesma tem a maior quantidade de registros dentro da base e foi escolhida com intuito de poder visualizar melhor as mudanças dentre as classes, nesta base a medição inicial foi aos 24 meses e a última foi aos 96 meses.

A parcela escolhida foi a de número 129 e a mesma continha classes de diâmetro de 1 cm até 23 cm, a Figura 40 apresenta a idade inicial onde cada cor é definida para as suas respectivas classes, nas seguintes Figuras são apresentadas as variações em relação as medições.

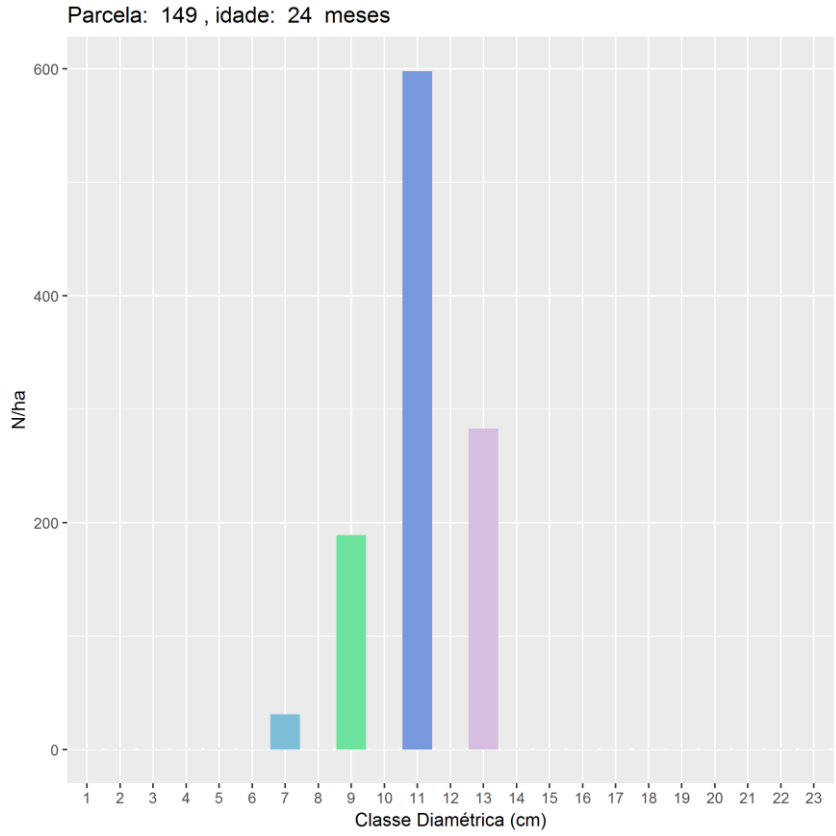


Figura 41 Incremento Diamétrico para parcela 149 aos 36 meses.

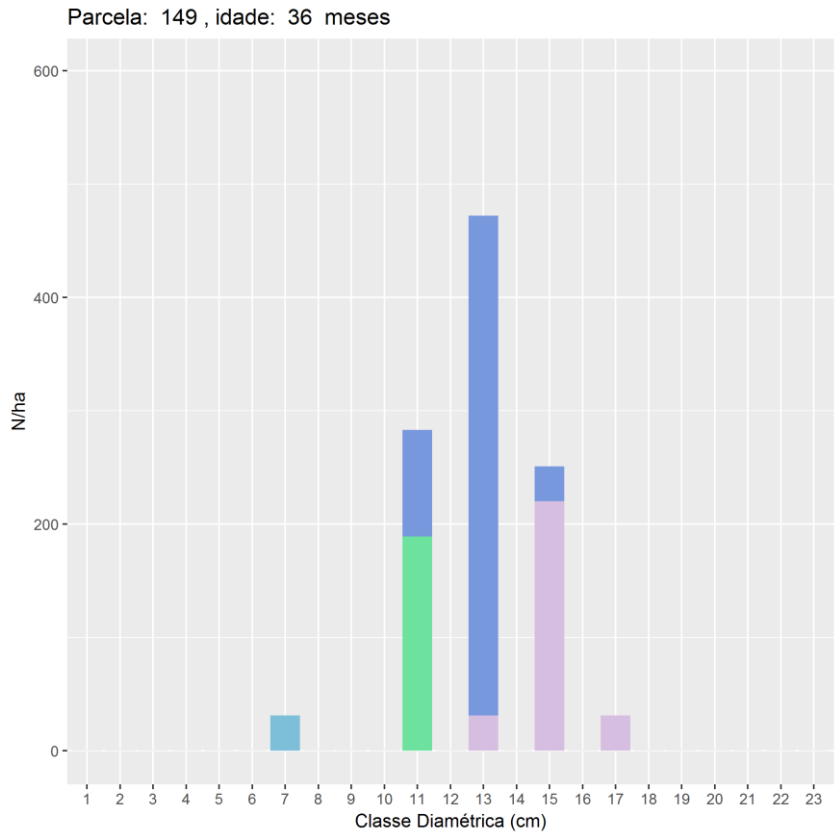


Figura 42 Incremento Diamétrico para parcela 149 aos 24 meses.

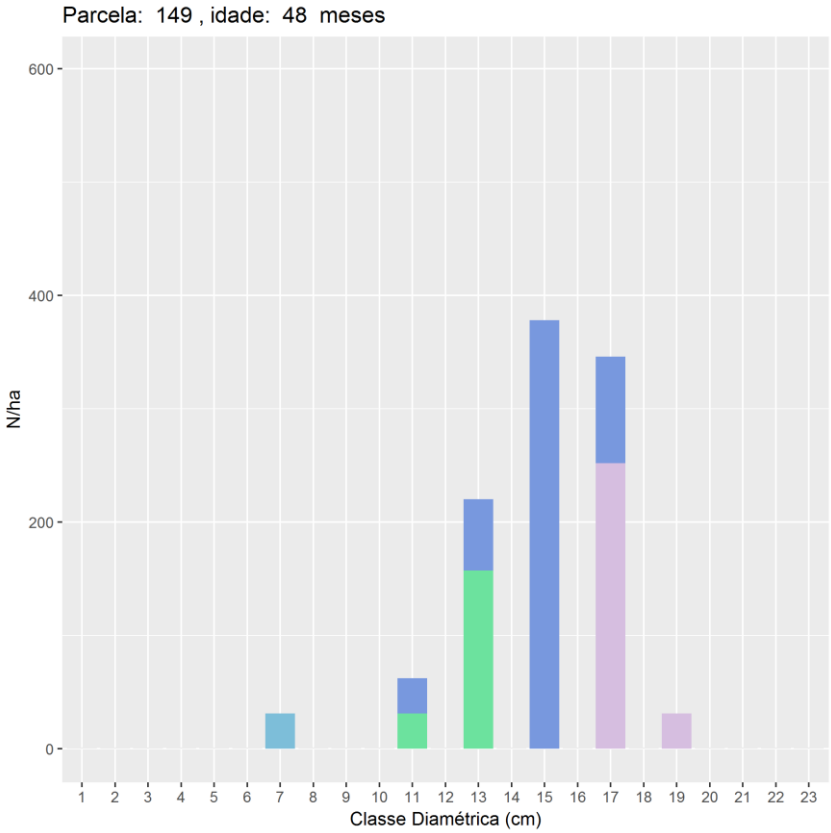


Figura 43 Incremento Diamétrico para parcela 149 aos 48 meses.

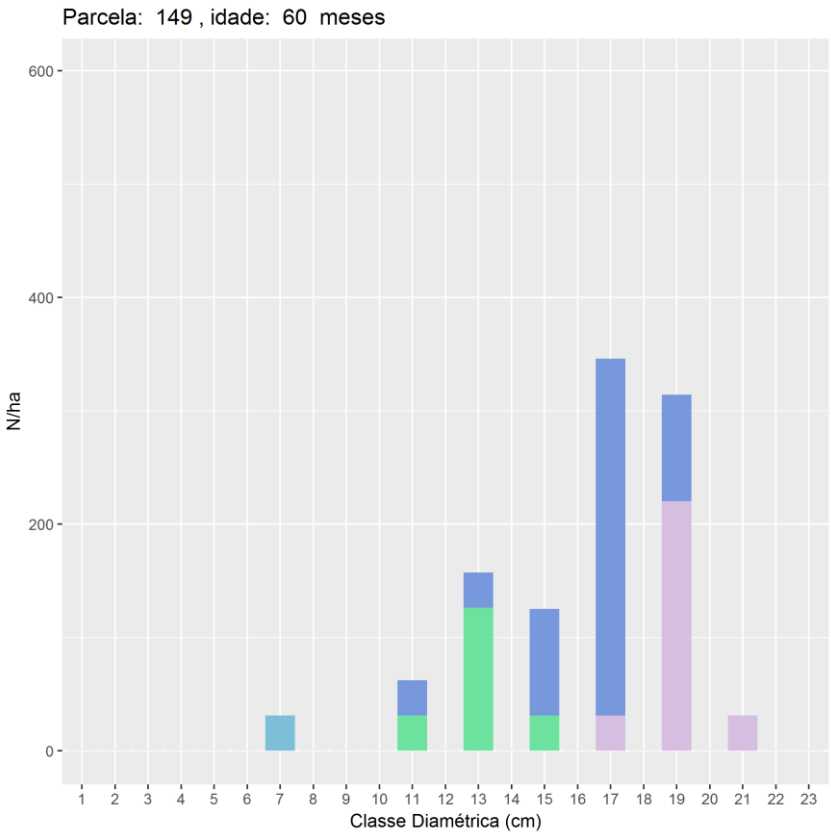


Figura 44 Incremento Diamétrico para parcela 149 aos 60 meses.

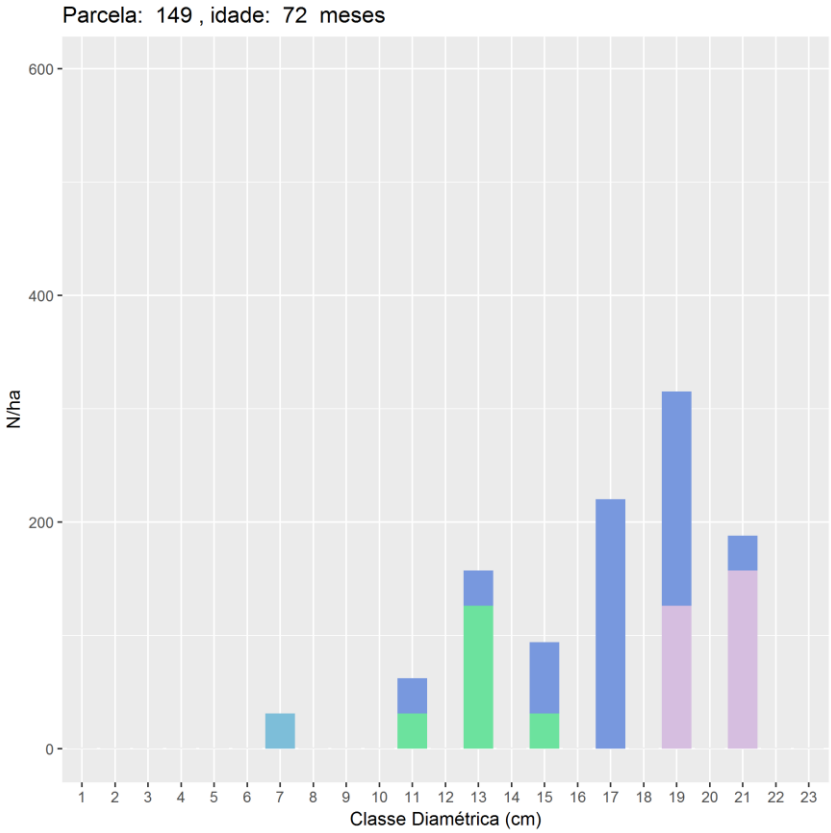


Figura 45 Incremento Diamétrico para parcela 149 aos 72 meses.

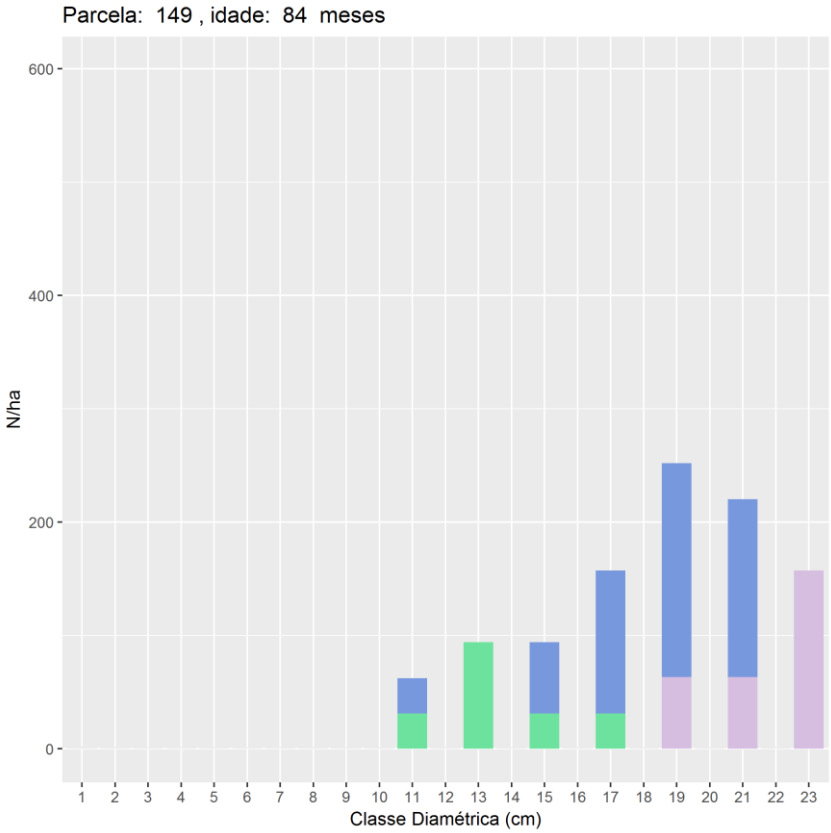


Figura 46 Incremento Diamétrico para parcela 149 aos 84 meses.

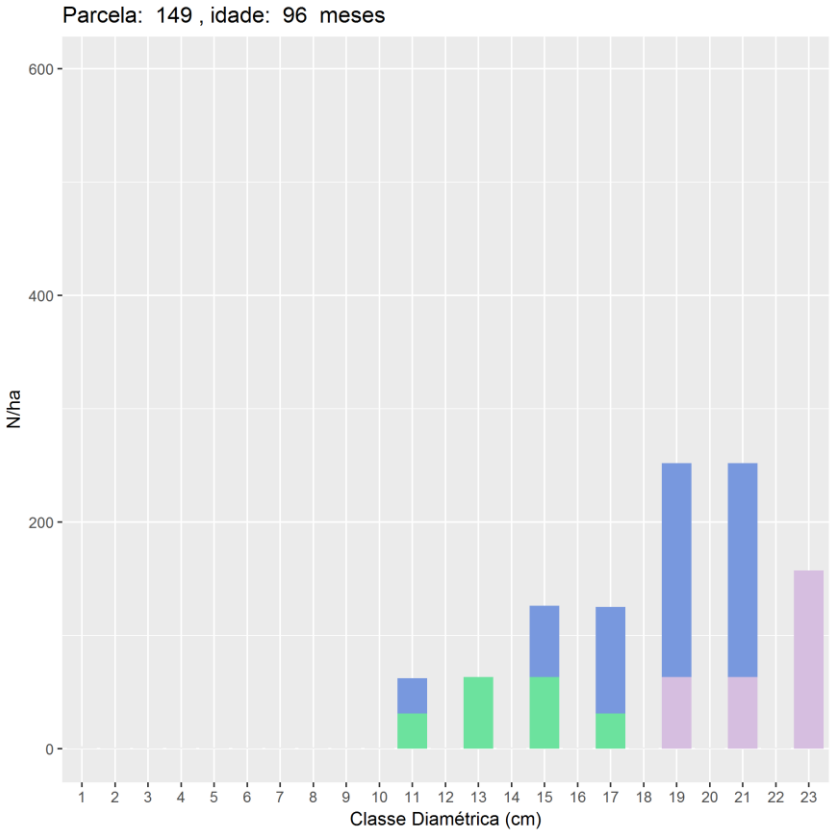


Figura 47 Incremento Diamétrico para parcela 149 aos 96 meses.

As Figuras 41,42,43,44,45,46,47 apresentam a dinâmica de mudança de classe para as idades 24,36,48,60,72,84,96 respectivamente, pode-se perceber que da idade 24 para 36 a principal mudança foi da classe 9 para classe 11 e da classe 11 para a classe 13, distanciando um pouco a idade se olharmos a idade 72 podemos ver mais claramente a separação e migração das classes de diâmetro para classes como 19 e 22.

Por fim, na última medição aos 96 meses, percebe-se que as árvores da classe a maioria das árvores que na primeira idade estavam na classe 13 chegaram a classe 23, ou seja, dentro de desse período de tempo a maioria das árvores cresceram 10 cm. Um exemplo de comportamento incomum visto mesmo na primeira idade é a classe 7 que se mostrou estagnada quanto ao crescimento nas idades 24 a 72 podendo ser fruto de uma análise mais aprofundada utilizando os outros métodos disponível no pacote afim de encontrar inconformidades quanto a qualidade dos dados presentes na parcela 149.

	cod_id	age_24	age_36	age_48	age_60	age_72	age_84	age_96
1	5_149_10_32__2_0	0	1	0	1	0	1	0
2	5_149_10_32__2_1	11	11	0	1	0	1	0
3	5_149_11_32__2_1	11	2	4	0	2	2	0
4	5_149_12_32__2_1	11	2	2	2	0	2	0
5	5_149_13_32__2_1	13	2	2	2	2	2	0
6	5_149_14_32__3_1	11	2	2	2	2	0	0
7	5_149_15_32__3_1	13	2	2	2	0	0	0
8	5_149_16_32__3_1	11	0	2	0	0	2	0
9	5_149_17_32__3_1	11	2	2	2	0	0	0
10	5_149_18_32__3_1	11	2	2	2	0	2	0

Figura 48 Saltos de classes da parcela 149.

A imagem 48 apresenta os 10 primeiros registros da tabela de saltos da parcela 149, pode-se perceber que na primeira idade (24) que existem registros que deram 11 saltos de 12 meses para 24 meses, ou seja, mudaram de classe 11 vezes, isso pode gerar dúvidas quanto a árvore podendo ser um dado inconsistente quanto a sua projeção de crescimento, gera a questão a seguinte questão: uma árvore pode realizar 11 saltos dentro de 12 meses? Cabe ao profissional que fará utilização do pacote averiguar essa ocorrência de saltos, portanto, isso serve como pontapé inicial para uma averiguação mais afunda quanto a parcela 149 em especial as árvores que apresentam esse tipo de comportamento.

5.3 Comparativo dos Estudos de Caso

O quadro 19 apresenta um comparativo dos dois estudos de caso realizados, o mesmo relaciona os métodos de detecção de inconsistências aplicados, os resultados obtidos e a dimensão de qualidade que foi afetada.

Método de Análise	**Resultados do 1º Estudo de Caso**	**Resultados do 2º Estudo de Caso**	**Dimensões de qualidade relacionadas**
check_measurement_ages (dap,ht)	7847	1433	Consistência
check_ages	765	322	Consistência
check_parcel_different_spacing	0	0	Consistência
check_clone_different_plot	0	0	Consistência
check_dead_state	267	0	Consistência
check_measurement_state	72886	1040	Consistência, Completude, Pontualidade
check_undefined_spacing	112	0	Completude, Pontualidade
check_zero_measurement	66072	0	Consistência, Completude, Pontualidade

Quadro 19 Comparativo dos estudos de caso realizados

Inicialmente temos o método *check_measurement_ages*, o mesmo foi aplicado a duas variáveis de medição (“dap”, ”ht”), no primeiro estudo de caso foram encontrados 7847 registros inconsistentes já no segundo foram encontrados 1433, a dimensão afetada foi a consistência pois todos os registros encontrados estavam inconsistentes a uma regra de negócio, por fim o método encontrou 9280 registros errados nos estudos de casos feitos. Já o método *check_ages* identificou 765 registros errados no primeiro estudo de caso e 322 no segundo, a dimensão afetada foi a consistência, e o método encontrou 1087 registros errados nos dois estudos de caso.

Os métodos *check_parcel_different_spacing*, *check_clone_different_plot* não identificaram nenhum registro errado, porém caso identificasse a dimensão afetada seria a consistência. Por outro lado, o método *check_dead_state* encontrou 267

registros errados no primeiro estudo de caso e zero registros no segundo, a dimensão afetada foi a de consistência. Quanto ao método *check_measurement_state* foram encontrados 72886 registros inconsistentes no primeiro estudo de caso realizado e no segundo 1040 totalizando 73926 registros encontrados por esse método, as dimensões afetas foram consistência, completude e pontualidade.

O método *check_undefined_spacing* identificou 112 registros errados no primeiro estudo de caso e no segundo zero, as dimensões afetadas foram a completude e a pontualidade uma vez que os registros não estavam disponíveis para uso muito menos completos. Por fim o método *check_zero_measurement* encontrou 66072 registros inconsistentes no primeiro estudo de caso e no segundo zero, as dimensões afetadas foram consistência, completude e pontualidade.

6 CONCLUSÃO

Neste trabalho foi desenvolvido um pacote de funções *Forest Data Quality* (FDQ) para realização de análises em bases de dados florestais, as mesmas objetivando a busca por inconsistências em base de dados de inventário florestal que comprometem a qualidade dos dados com foco nas dimensões de consistência, completude e pontualidade.

O pacote conta com 8 métodos de análises baseados nas medidas de consistência propostas por Fraga Filho (2016) e com o processo de incremento diamétrico proposto por Demolinari (2006), atingindo desta forma o objetivo do desenvolvimento de uma ferramenta capaz de avaliar a qualidade de dados de bases florestais. Com o mesmo foi possível avaliar a qualidade dos dados utilizados como entrada em modelos de crescimento e produção que são os dados de inventário e além identificar inconsistências quanto a qualidade dos dados.

Como constatado pelos dois estudos de caso realizados, o pacote se mostrou eficiente na identificação de inconsistências nas bases de dados analisadas, atingindo dessa forma as hipóteses esperadas sobre o trabalho. Uma vez que pacote já está publicado no repositório oficial de pacotes do R (CRAN) e possui acesso livre e gratuito para todos, espera-se que o pacote seja utilizado por estudantes e profissionais para receber sugestões, feedbacks afim de melhorar eventuais falhas para constante evolução. Além disso, com sua documentação disponibilizada em inglês espera-se que o conhecimento sobre *Forest Data Quality* seja difundido por inúmeros países afim de torna-lo mais conhecido e utilizado.

6.1 Trabalhos Futuros

Uma vez que o FDQ é um pacote de código livre e gratuito disponibilizado no repositório central do R (CRAN), espera-se que o mesmo seja evoluído através da adição de novos métodos de análise ou mesmo a melhoria dos já existentes, fazendo assim com que o mesmo se torne mais completo e possa atender uma gama maior de necessidades da área florestal. Além disso, espera-se também que o pacote seja utilizado na elaboração de novos pacotes assim como utilizou-se pacotes para elaborar o FDQ.

Sugere-se ainda que seja realizada a implementação de um software que atue como uma interface gráfica para o pacote e proporcione um ambiente de gerenciamento de configuração de suas bases atuando em cima do versionamento e controle de mudança das mudanças decorridas durante a realização das análises, toda documentação como casos de uso, modelos de domínio, diagramas do software encontra-se no apêndice B deste trabalho.

Por fim, espera-se que outros estudos de caso sejam realizados com dados provenientes de outras amostras de dados florestais a fim de comprovar mais ainda a eficácia do pacote na identificação de erros relacionados a qualidade de dados de bases florestais.

7 REFERÊNCIAS BIBLIOGRÁFICAS

AGUIRRE, L. A. **Introdução a identificação de sistemas**: Técnicas lineares e não-lineares aplicadas a sistemas reais. 3 ed. Belo Horizonte: UFMG, 2007. 729 p.

AMMAR, R. randomcoloR: Generate Attractive Random Colors. Disponível em:< https://CRAN.R-project.org/package=randomcoloR >. Acesso em: 14/11/2016.

ANDRIOTTI, J. L. S. **Técnicas estatísticas aplicáveis a tratamento de informações oriundas de procedimentos laboratoriais**. Porto Alegre, 2005.

ARAÚJO, T.M.E.; BATISTA, S.L.M. **Uma visão sobre a qualidade dos dados,** 2007, disponível em: <http://www.devmedia.com.br/uma-visao-sobre-a-qualidade-dos-dados/6973>. Último acesso em 12/05/2016.

AUGUSTO, C.; FINGER, G.; NICOLETTI, M. F.; et al. Exatidão de dendrômetros ópticos para determinação do volume de árvores em pé. **Ciência Florestal**, v. 25, n. 2, p. 395–404, 2015.

ASSUNÇÃO, F. **Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos.** 2012. 74 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2012.

BARBIERI, C. **Uma visão sintética e comentada do Data Management Body of Knowledge (DMBOK)**. Fumsoft - Belo Horizonte, 2013.

BATINI, C.; SCANNAPIECA, M. **Data Quality: Concepts, Methodologies and Techniques**. Roma: Springer-Verlag Berlin Heidelberg, 2006.

BENTANCOURT, S. S. M. P.; ROCHA, R. P. Metadados de qualidade e visibilidade na comunicação científica. **Revista eletrônica de biblioteconomia e ciência da informação,** Florianópolis, v. 17, n. 2, p. 82-101, 2012.

BURKHART, H. E. Stand modelling for radiata pine in New Zealand. **New Zealand Journal of Forestry**, 1977. Disponível em: <http://agris.fao.org/agris-search/search.do?recordID=NZ19800496000>. Acesso em: 02/06/2016.

CAMPOS, A. L. de S. **Indicadores de Qualidade de Dados em Data Warehouse. Monografiade conclusão de curso. UFJF, Juiz de Fora, 2005. 51.**

CAMPOS, J.C.C.; LEITE, H.G. **Mensuração florestal:** perguntas e respostas. 2. Ed Viçosa: Editora UFV, 2006. 470 p.

CAMPOS, J.C.C.; LEITE, H.G. **Mensuração florestal:** perguntas e respostas. 4. Ed Viçosa: Editora UFV, 2013. 605 p.

CARNEIRO, A. C. M. S. **Análise de dados longitudinais através de modelos marginais.** 2012. 50 p. Trabalho de conclusão de curso (Dados longitudinais) , Juiz de Fora.

CEPEA. Centro de Estudos Avançados em Economia Aplicada, 2015. Disponível em < http://cepea.esalq.usp.br/pib/>. Acesso em: 25 jun.2016.

CRAN, The Comprehensive R Archive Network. Disponível em: <https://cran.r-project.org/index.html>. Acesso em 14/11/2016.

Data Quality Assessment Framework (DQAF), 2012.

DQAF- Generic Framework, 2003.

DEMOLINARI, R.A. **Crescimento de povoamento de eucalipto não-debastados**, 2006. Tese (Magister Scientiae em Ciências Florestais) – Ciência Florestal, Universidade Federal de Viçosa, Viçosa, 2006. Dissertação (Mestrado em Engenharia Florestal) – Universidade Federal de Viçosa, Viçosa, 2014.

DOWLE, M.; SHORT, T.; LIANOGLOU, S.; SRINIVASAN, A. data.table: Extension of data.frame. Disponível em: < https://cran.r-project.org/package=data.tablel>. Acesso em: 14/11/2016

DUVEMO, K. **The influence of data uncertainty on planning and decision processes in forest management,** 2009.

ENGLISH, L. P. 1999. **Improving Data Warehouse and Business Information Quality**, John Wiley & Sons, Inc.

FILHO, C. V. F. **Forest Measure**. Disponível em: <http://sourceforge.net/projects/forestmeasure/>. Acesso em: 17 jun. 2016, 2014.

FILHO, C. V. F; SIMIQUELI, A.P;SILVA, G.F; ALTOE, W.A.S. **Fgmutils:** Forest Grow Model Utilities. . [S.l: s.n.]. Disponível em: < https://cran.r-project.org/web/packages/Fgmutils/>. Acesso em: 17 jun. 2016, 2015.

FILHO, J.T. Transformando dados corporativos em inteligência competitiva, 1999. Disponível em: <http://crdpublicidade.com.br/files/artigo131.pdf>. Acesso em: 25 jun. 2016.

FRAGA FILHO, C. V. **Modelagem para prognose do crescimento e produção de eucalipto em diferentes níveis de abordagem.** 2016. 281 f. Tese (Doutorado) - Centro de Ciências Florestais, Universidade Federal do Espírito Santo, Jerônimo Monteiro, 2016.

GERTNER, G. Prediction bias and response surface curvature. **Forest science**, v. 37, n. 3, p. 755–765, 1991. Society of American Foresters.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining**: um guia pratico: conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005. 261 p.

GROTHENDIECK, G. sqldf: Perform SQL Selects on R Data Frames. Disponível em:< https://cran.r-project.org/web/packages/sqldf/index.html>. Acesso em: 14/11/2016

GUJARATI, D. N. **Econometria Básica.** 3. ed. Retirar: Bookman, 2000.

HUANG, K., LEE, Y. and WANG, R. 1999, **Quality Information and Knowledge. Prentice Hall**, Upper Saddle River: N.J.

HUOT, R. **Métodos Quantitativos para as Ciências Humanas.** Lisboa: Instituto Piaget, 2002. 384 p.

INMON, W. H. **Como Construir o Data Warehouse**. Rio de Janeiro: Campus, 1997.

ISLAM, M. N.; KURTTILA, M.; MEHTÄTALO, L.; HAARA, A. Analyzing the effects of inventory errors on holding-level forest plans: the case of measurement error in the basal area of the dominated tree species. **Silva Fennica**, v. 43, n. 1, p. 71–85, 2009.

Finnish Society of Forest Science.

ISO/IEC 25024 Software Engineering: Software product Quality Requirements and Evaluation (SQuaRE) Measurement of quality in use, 2015.

KAHN, B. K., STRONG, D. M. 1998. Product and Service Performance Model for Information Quality: An Update. **in Proceedings of the 1998 Conference on Information Quality**. Cambridge, MA: pp. 102-115.

KANGAS, A; MALTAMO, M. **Forest Inventory: Methods and Applications.**

LEITE, H. G. **Ajuste de um modelo de estimação de freqüência e produção por classe de diâmetro, para povoamentos de Eucalyptus saligna Smith**, 1990. 97f. Dissertação (Mestrado em Ciência Florestal) – Universidade Federal de Viçosa, Viçosa, 1990.

LIMA, C.R. DE A. **Gestão da qualidade dos dados e informações dos sistemas de informação em saúde: subsídios para a construção de uma metodologia adequada ao Brasil,** 2010. Escola Nacional de Saúde Pública Sérgio Arouca.

LOSHIN, D. **Enterprise Knowledge Management:** the data quality approach. 2001. 493p.

MAGALHÃES, J. G. DE S. **Prognose e Regulação de Florestas de eucalipto utilizando modelo de crescimento e inteligência computacional**, 2014. 66f.

MAITA, A. R. C. **Um estudo da aplicação de técnicas de inteligência computacional e de aprendizado de máquina em mineração de processos de negócio.** 2016. 130 f. Dissertação (Mestrado) - Universidade de São Paulo, São Paulo, 2016.

MALAVERRI, J. E. G; MEDEIROS, C. B. Data Quality in Agriculture Applications**. Proceedings XIII GEOINFO**, Campos Do Jordão, Brasil, p. 128-139,2012.

MARCATTI, G. E. **Caminhamento ótimo para acesso à parcelas de inventário florestal**, 2013. 41f. Dissertação (Ciências Florestais) – Universidade Federal de Viçosa, Viçosa, 2013.

MORAIS, C.M. **Escalas de Medida, Estatística Descritiva e Inferência Estatística.** Bragança, 2005.

MONTGOMERY, D.C. (2001) - **Introduction to Statistical Quality Control**, 4th edition, John Wiley and Sons.

PIENAAR, L. V.; SHIVER, B. D. Survival functions for site prepared slash pine plantations in the flat woods of Georgia Northern Florida. **Southern Journal Forestry 5**, p.59-62, 1981.

PIETILÄ, I.; KANGAS, A.; MÄKINEN, A.; MEHTÄTALO, L. Influence of growth prediction errors on the expected losses from forest decisions. **Silva Fennica**, v. 44, n. 5, p. 829–843, 2010.

R DEVELOPMENT CORE TEAM, R. D. C. **R: A language and environment for statistical computing**. **R foundation for Statistical Computing**. Vienna, Austria: sn, 2005.

Reis, E. **Estatística Descritiva**. Lisboa, 1996.

RIBEIRO, A. P. T. **Cenário de gestão de dados no serpro comparado à referência conceitual do guia DMBOK.** Porto Alegre, 2014.

ROMANIUK, D. S. **Modelagem do crescimento e da produção em plantios de *Pinus taeda* L. na região Centro-Sul do Paraná** 2015. 71 f. Dissertação (Mestrado) – Departamento de Engenharia Florestal, Setor de Ciências Agrárias e Ambientais, Universidade Estadual do Centro-Oeste, Irati - PR, 2015.

ROMUALDO, R. M. A. **Sistema para estimativa de parâmetros de funções de probabilidades para variáveis contínuas com o método da máxima verossimilhança.** 2015. 81p. Universidade Federal do Espírito Santo, Alegre.

SANQUETTA, C. R. et al**. Inventários Florestais: Planejamento e Execução.**

SCOLFORO, J. R. S. **Biometria florestal: modelos de crescimento e produção florestal**. Lavras, 2006.

SEBASTIAN-COLEMAN, L. **Measuring data quality for ongoing improvement: a data quality assessment framework**. Waltham: Elsevier, 2013.

SOARES, C. P. B.; PAULA NETO, F.; SOUZA, A. L. **Dendrometria e inventário florestal.** Viçosa, MG: Universidade Federal de Viçosa, 2006. 276p.

SOARES, J. D. R. **Modelagem Matemática Na Previsão da Colheita de Bananeira:** Regressão Linear Múltipla X Redes Neurais Artificiais (Doutorado em Fitotecnia) – Universidade Federal de Lavras, Lavras, 2010.

SOUZA, A.L.; ARAUJO, P.A.; CAMPOS, J.C.C. e PAULA NETO, F. Dinâmica de crescimento em diâmetro de uma floresta primária sem interferência: uma análise pelo tempo de passagem entre classes diamétricas. **Revista Árvore**, Viçosa, MG: v.17, n.2, p.129-145, 1993.

VANCLAY, J. K. Modelling forest growth and yield: applications to mixed tropical forests. **School of Environmental Science and Management Papers**, p. 537, 1994.

VEIGA, R. A. DE A. **Dendrometria e inventário florestal**. Botucatu, FEPAF, 1984.

WEISBERG, Sanford. **Applied Linear Regression.** 3. ed. Minneapolis, Minnesota, 2005. 330 p.

WICKHAM, H. The split-apply-combine strategy for data analysis. **Journal of Statistical Software**, v. 40, n. 1, p. 1–29, 2011. Disponível em: <Http://www.jstatsoft.org/index.php/jss/article/view/v040i01/v40i01.pdf>. Acesso em: 14/11/2016.

WICKHAM, H.; CHANG, W. ggplot2: Creating Elegant Data Visualisations Using the Grammar of Graphics. Disponível em:<https://CRAN.R-project.org/package=ggplot2>. Acesso em: 14/11/2016.

8 APÊNDICES

8.1 APÊNDICE A – Scripts De Execução Dos Métodos de Análises

Método: check_measurement_ages

```
require(fdq)
dfResultado = check_measurement_ages(dfBrutosPareado,"dap1","dap2")
summary(dfResultado)
head(dfResultado)
dfResultado = check_measurement_ages(dfBrutosPareado,"ht1","ht2")
summary(dfResultado)
head(dfResultado)
```

Método: check_ages

```
require(fdq)
dfResultado = check_ages(dfBrutosPareado,rounded_age1 =
"idadearred1",rounded_age2 = "idadearred2",months = TRUE)
summary(dfResultado)
head(dfResultado)

dfResultado = check_ages(dfBrutosPareado,rounded_age1 =
"idadearred2",rounded_age2 = "idadearred1",months = TRUE)
summary(dfResultado)
head(dfResultado)
```

Método: check_parcel_different_spacing

```
require(fdq)
dfResultado =
check_parcel_different_spacing(dfBrutos,"parcela","espacamento1","espacamen
to2",c("parcela"))
summary(dfResultado)
head(dfResultado)

dfResultado =
check_parcel_different_spacing(dfDados,"parcela","espacamento1","espacament
o2",c("parcela"))
summary(dfResultado)
head(dfResultado)
```

Método: check_clone_different_plot

```
require(fdq)
dfResultado =
check_clones_different_plot(dfBrutos,"parcela","clone",c("parcela","clone")
)
summary(dfResultado)
head(dfResultado)

dfResultado =
check_clones_different_plot(dfDados,"parcela","clone",c("parcela","clone"))
summary(dfResultado)
head(dfResultado)
```

Método: check_dead_state

```
require(fdq)
dfResultado = check_dead_state(dfBrutos,"estado",c("dap","ht"))
summary(dfResultado)
head(dfResultado)

dfResultado = check_dead_state(dfDados,"estado",c("dap","ht"))
summary(dfResultado)
head(dfResultado)
```

Método: check_measurement_state

```
require(fdq)
dfResultado = check_measurements_state(dfBrutos,c("dap","ht"),"estado")
summary(dfResultado)
head(dfResultado)

dfResultado = check_measurements_state(dfDados,c("dap","ht"),"estado")
summary(dfResultado)
head(dfResultado)
```

Método: check_undefined_spacing

```
require(fdq)
dfResultado = check_undefined_spacing(dfBrutos,"espacamento")
```

```
summary(dfResultado)
head(dfResultado)

dfResultado = check_undefined_spacing(dfDados,"espacamento")
summary(dfResultado)
head(dfResultado)
```

Método: check_zero_measurement

```
require(fdq)
dfResultado = check_zero_measurement(dfBrutos,c("dap","ht"),"estado","N")
summary(dfResultado)
head(dfResultado)

dfResultado = check_zero_measurement(dfDados,c("dap","ht"),"estado","N")
summary(dfResultado)
head(dfResultado)
```

Método: check_diametric_increase

```
require(fdq)
check_diametric_increase(dfBrutos,"cod_id","parcela",356,"idade","idadearre
d","classedediametro",dfBrutosNHA,"parcela","classedediametro","idadearred"
,"nclasseha")
plot(p356i12)
```

Método que calcula os saltos por classe de diâmetro

```
require(fdq)
dfSaltos =
calculate_jumps_ages(dfBrutos,"cod_id","idadearred","classedediametro")
summary(dfSaltos)
head(dfSaltos)

check_diametric_increase(dfDados,"cod_id","parcela",149,"idade","idadearred
","classedediametro",dfDadosNHA,"parcela","classedediametro","idadearred","
nclasseha")
plot(p149i12)
```

Método que calcula os saltos por classe de diâmetro

```
require(fdq)
dfSaltos =
calculate_jumps_ages(dfDados,"cod_id","idadearred","classedediametro")
summary(dfSaltos)
head(dfSaltos)
```

8.1.1 Script realizado para obter a interseção de resultados

```
dfIntersecao = merge(dfResultado1,dfResultado2,by=c("cod_id"))

summary(dfIntersecao)
head(dfIntersecao)
```

	cod_id	ano_medicao1.x	ano_medicao2.x	dap1.x	dap2.x	ht1.x	ht2.x	espacamento1.x	espacamento2.x
1	16_743_10_138__2_1	2006	2009	10.0	0.0	0.0	0.00	3,3 x 3	3,3 x 3
2	23_895_22_826__4_1	2007	2009	8.0	0.0	0.0	0.00	3,3 x 3	3,3 x 3
3	25_740_12_123__2_1	2006	2009	10.4	0.0	0.0	0.00	3,3 x 3	3,3 x 3
4	3_333_29_24__3_1	2006	2008	21.6	22.9	32.2	0.00	3,3 x 3	3,3 x 3
5	3_333_47_24__5_1	2006	2008	12.8	12.6	20.9	22.30	3,3 x 3	3,3 x 3
6	30_772_3_269__1_1	2006	2008	9.0	0.0	0.0	0.00	3,3 x 3	3,3 x 3
7	30_773_5_269__1_1	2006	2008	10.0	0.0	10.3	0.00	3,3 x 3	3,3 x 3
8	30_773_5_269__1_1	2006	2008	10.0	0.0	10.3	0.00	3,3 x 3	3,3 x 3
9	30_780_24_262__4_1	2006	2008	8.1	0.0	11.3	0.00	3,3 x 3	3,3 x 3
10	30_780_24_262__4_1	2006	2008	8.1	0.0	11.3	0.00	3,3 x 3	3,3 x 3

Figura 49 Exemplo de base resultado após interseção

8.2 APÊNDICE B – Documentação do Pacote *Forest Data Quality*

Package 'fdq'

December 11, 2016

Type Package

Title Forest Data Quality

Date 2016-12-11

Version 0.2

Maintainer Caíque de Oliveira de Souza <forestdataqualitysoftware@gmail.com>

Description

Contains methods of analysis of forest databases, the purpose of the analyzes is to evaluate the quality of the data present in the databases focusing on the dimensions of consistency, punctuality
and completeness. Databases can range from forest inventory data to growth model data. The package has
methods to work with large volumes of data quickly, in addition in certain analyzes it is possible to generate the graphs for a better understanding of the analysis and reporting of the analyzed analysis.

License GPL-3

Encoding UTF-8

LazyData true

Depends R(>= 3.0), Fgmutils,data.table

Imports ggplot2, sqldf, randomcoloR, plyr, utils, stats

RoxygenNote 5.0.1

NeedsCompilation no

Author Caíque de Oliveira de Souza [aut, cre],
Clayton Vieira Fraga Filho [ctb, dtc],
Miquéias Fernandes [ctb]

Repository CRAN

Date/Publication 2016-12-11 20:35:12

R topics documented:

calculate_jumps_ages 2
check_ages 3

1

check_clones_different_plot . . . 3
check_dead_state . . . 4
check_diametric_increase . . . 4
check_existing_ages . . . 5
check_existing_place . . . 6
check_existing_plots . . . 6
check_measurements_state . . . 7
check_measurement_ages . . . 7
check_parcel_different_spacing . . . 8
check_size_age_plot . . . 8
check_undefined_spacing . . . 9
check_variables . . . 9
check_zero_measurement . . . 10
correct_probability . . . 10
find_missing_age . . . 11
find_missing_place . . . 11
find_missing_variable . . . 12
generate_colors_diameter_class . . . 12
generate_diameter_classes . . . 13
generate_initial_diameter_class . . . 13
generate_limiar . . . 14
generate_new_color . . . 14
generate_number_hectare . . . 15
generate_number_hectare_class . . . 15
get_ages . . . 16
get_max . . . 16
get_min . . . 17
get_place . . . 17
mount_list . . . 18
mount_query . . . 18
plot_cols . . . 19
sort_columns_crescent . . . 19

Index **20**

`calculate_jumps_ages` *calculate_jumps_ages*

Description

This function generates the jumps per diameter class among the ages in the database

Usage

```
calculate_jumps_ages(database, id_field, rounded_ages_field,
  diameter_class_field)
```

Arguments

database	data.frame data.table
id_field	string name of column with unique identifier of each tree
rounded_ages_field	string name of column with rounded ages
diameter_class_field	string name of column with diametric classes

---

check_ages *check_ages*

Description

This analysis verifies age differences on a paired basis, if the rounded ages are in months the check is if the difference is 12 months, if it is in year the consecutive ages should only present difference of 1 year, doubts about how to pair your base consult The Fgmutils package

Usage

```
check_ages(data_base, rounded_age1, rounded_age2, months = FALSE)
```

Arguments

data_base	data.frame data.table
rounded_age1	string name of column rounde age one
rounded_age2	string name of column rounde age two
months	TRUE for age in months or FALSE for age in years

---

check_clones_different_plot *check_clones_different_plot*

Description

This function checks if the clones of a tree have different plots

Usage

```
check_clones_different_plot(database, parcel_name, clone_name,
  variables_to_group)
```

Arguments

`database`	data.frame, data.table or any database
`parcel_name`	string name of the field containing the parcels
`clone_name`	string name of the field containing the clones
`variables_to_group`	string(s) variable (s) that you want to group the result of the analysis

---

`check_dead_state` *check_dead_state*

Description

This function checks if the base state field is equal to dead (M) and there is some kind of measurement

Usage

```
check_dead_state(data_base, state, measurement_variables)
```

Arguments

`data_base`	data.frame data.table or any database
`state`	string field name representing state column in database
`measurement_variables`	string vector that contains a set of measurement variables to be analyzed, this variables are names of columns in database

---

`check_diametric_increase` *check_diametric_increase*

Description

This function is an implementation of the process proposed by Demolinari (2006) in his master's thesis that is the diametric increase, given a plot it is possible to follow all the growth of the trees contained in the same at all ages present in the plot, this is done through Of colored color graphics where each color represents a class of diameter and a meditation that as an advance for each age is generated a new graphic that contains as changes represented by cores or a quantity of trees that migrate of class, this can have an identification Of inconsistent behaviors within the plot.

Usage

```
check_diametric_increase(database, id_field, plot_field, plot_analyzed,
  ages_field, rounded_ages_field, diameter_classes_field, database_nhaclasses,
  plot_field_database_nhaclassses, diameter_classes_field_database_nhaclasses,
  rounded_ages_field_database_nhaclasses, nha_classes_database_nhaclasses)
```

Arguments

`database,` data.table, data.frame or any database

`id_field,` string with the name of column containing the unique identifier of each tree

`plot_field,` string with the name of column containing the plots of database

`plot_analyzed,` the number of plot to be analyzed, example: 356

`ages_field,` string with the name of column containing the ages of each tree

`rounded_ages_field,` string with the name of column containing the rounded ages of each tree, example: 12, 36

`diameter_classes_field,` string with the name of column containing the diameter class of each tree

`database_nhaclasses,` data.table, data.frame, or any database containg the NHa Class of database used in this analysis. To obtain this database you simply have to submit the original base the generate_number_hectare_class function contained in this package

`plot_field_database_nhaclassses,` string with the name of column containing the plots of database resulting from the generate_number_hectare_class function

`diameter_classes_field_database_nhaclasses,` string with the name of column containing the diameter class of each tree of database resulting from the generate_number_hectare_class function

`rounded_ages_field_database_nhaclasses,` string with the name of column containing the rounded ages of each tree of database resulting from the generate_number_hectare_class function

`nha_classes_database_nhaclasses,` string with the name of column containing the NHa Class of database resulting from the generate_number_hectare_class function

---

`check_existing_ages` *check_existing_ages*

Description

This function checks if a given set of ages exists in a database column

Usage

```
check_existing_ages(database, ages_name, ages_to_check)
```

Arguments

database	data.frame data.table or any database
ages_name	string name of the column representing ages
ages_to_check	string name/vector of the column (s) representing ages to be checked

---

check_existing_place *check_existing_place*

Description

This function checks whether a particular set of sites or locations exists in a database column

Usage

```
check_existing_place(database, place_name, places_to_check)
```

Arguments

database	data.frame, data.table or any database
place_name	string name of the column representing site or place
places_to_check	value(s) to be checked, example: c(12,21,33)

---

check_existing_plots *check_existing_plots*

Description

This function checks if a particular set of parcels exists in a database column

Usage

```
check_existing_plots(database, plots_name, plots_to_check)
```

Arguments

database	data.frame, data.table or any database
plots_name	string column name representing parcels in the base
plots_to_check	value(s) to be checked, example: c(356,122)

check_measurements_state

check_measurements_state

Description

This function checks if there is a measurement variable with value equal to 0 and if the respective states are different from M, F, A

Usage

```
check_measurements_state(data_base, measurement_variables, state)
```

Arguments

data_base	data.frame, data.table or any database
measurement_variables	set of variables to be analyzed, this set can be a vector of string with names of colunms
state	string name of the field that represents the state in database

check_measurement_ages

check_measurement_ages

Description

This function verifies if measurement variables have records of type DAP2 <DAP1, HT2 <HT1 in consecutive ages i + 1 and i tt is necessary that the base is already paired to perform such analysis, to know more about pairing consult the Fgmutils package

Usage

```
check_measurement_ages(data_base, measurement_variable1, measurement_variable2)
```

Arguments

data_base	data.frame, data.table or any database
measurement_variable1	string field containing the measurement variables at age 1
measurement_variable2	string field containing the measurement variables at age 2

check_parcel_different_spacing *check_parcel_different_spacing*

Description

This function checks for partitions with different spacing at i and i + 1 ages, it is necessary that the base be paired including the field representing the spacing, doubts about how to pair its base see the Fgmutils package

Usage

```
check_parcel_different_spacing(database, parcel_name, spacing_age1,
  spacing_age2, variables_to_group)
```

Arguments

database	data.frame, data.table or any database
parcel_name	string containing the field name parcels in database
spacing_age1	string containing the name of the field spacing in the first age
spacing_age2	string containing the name of the field spacing in the second age
variables_to_group	variable (s) that you want to group the result of the analysis, this can be a vector os strings or strign name to group

check_size_age_plot *check_size_age_plot*

Description

This function checks if the age field is more than one age, returning TRUE to for yes and FALSE for no

Usage

```
check_size_age_plot(database, age_name)
```

Arguments

database	data.frame, data.table or any database
age_name	string containing the name of the column that represents age

check_undefined_spacing *check_undefined_spacing*

Description

This function checks if there is any record with undefined spacing (0 or NA)

Usage

```
check_undefined_spacing(data_base, spacings)
```

Arguments

data_base	data.frame, data.table or any database
spacings	string vector containing the name of the variable (s) than represent spacings in database

check_variables *check_variables*

Description

This function checks if the entered column exists within the base

Usage

```
check_variables(database, variables)
```

Arguments

database	data.frame, data.table or any database
variables	vector of strings with names of columns

Value

TRUE for all variables in database, or FALSE for variables not present in columns

Examples

```
test <- data.frame("tree","diametrer","N")
check_variables(test,c("tree","diameter"))
```

check_zero_measurement *check_zero_measurement*

Description

This analysis verifies which measurement variables have values equal to 0 and then checks if there are variables in the states that the user reported

Usage

```
check_zero_measurement(data_base, measurement_variables, state_name,
  states_to_check)
```

Arguments

data_base	data.frame, data.table or any database
measurement_variables	string vector containing name of the field(s) it represents measurement variable(s) to be analyzed
state_name	string vector containing the name of the variable than represents state in database
states_to_check	string vector containing the name of the the states to be checked, the user can inform this names in a string vector like ("F","N")

correct_probability *correct_probability*

Description

This function will adjust the probability after generation of surviving tree classes per hectare

Usage

```
correct_probability(database, plot_field, rounded_age_field, probability_field,
  nha_class_field)
```

Arguments

database	data.frame, data.table or any database
plot_field	string name of column with plots of database
rounded_age_field	string name of column with rounded ages, examples: 12,24,36,48

`probability_field`
: string name of column with probabilitys

`nha_class_field`
: string name of column with NHa (surviving tree classes per hectare) classes,

---

`find_missing_age` *find_missing_age*

Description

This function identifies the missing age values in the database and notifies them to the user.

Usage

```
find_missing_age(database, age_name, ages_to_check)
```

Arguments

`database`	data.frame, data.table or any database
`age_name`	string that contains the field name that represents age in database
`ages_to_check`	vector containing the values of ages to be checked like c(12,23,48)

`find_missing_place` *find_missing_place*

Description

This function identifies values of sites or locations in the database and notifies them to the user

Usage

```
find_missing_place(database, place_name, places_to_check)
```

Arguments

`database`	data.frame, data.table or any database
`place_name`	string that contains the field name representing site or place in database
`places_to_check`	vector containing the values of places/sites to be checked like c(21,33,48)

find_missing_variable *find_missing_variable*

Description

This function identifies non-existent column names in the database and informs the user

Usage

```
find_missing_variable(data_base, variables)
```

Arguments

data_base	data.frame, data.table or any database
variables	vector string that contains the name(s) of columns to be checked in database

generate_colors_diameter_class *generate_colors_diameter_class*

Description

This function generates a new random color for each diameter class in the base

Usage

```
generate_colors_diameter_class(database, diameter_classe_name)
```

Arguments

database	data.frame, data.table or any database
diameter_classe_name	string with the name of field (column) containing the diameter classes

generate_diameter_classes

generate_diameter_classes

Description

This function identifies non-existent column names in the database and informs the user

Usage

```
generate_diameter_classes(database, diameter_names, amplitude,
  name_of_diameter_class)
```

Arguments

database	data.frame, data.table or any database
diameter_names	string with name of the field that contains the diameters of database
amplitude	desired amplitude for class creation, example: 1,2,4,6,7
name_of_diameter_class	string with name you want for the field class of diameter

generate_initial_diameter_class

generate_initial_diameter_class

Description

This function generates the initial class field to aid in the process of diametric increasing

Usage

```
generate_initial_diameter_class(database, plot_name, age_name)
```

Arguments

database	data.frame, data.table or any database
plot_name	string with the name of field representing plots in database
age_name	string with the name of field representing rounded age

generate_limiar *generate_limiar*

Description

This function generates a threshold for certain values based on the field entered and a maximum value for setting

Usage

```
generate_limiar(database, field, max)
```

Arguments

database,	data.table, data.frame or any database
field,	string name with column you want to work on
max,	number with the maximum value you want to establish the threshold

generate_new_color *generate_new_color*

Description

This function generates a new random color without repeating the ones that were entered in the last field as parameter

Usage

```
generate_new_color(colors)
```

Arguments

colors	vector of strings containing existing colors, exemple: c("#6140bc" "#e75bf7" "#d15102" "#6a0b9e" "#e8ad4e")

generate_number_hectare

generate_number_hectare

Description

This function generates the NHa, field that represents the number of surviving trees per hectare

Usage

```
generate_number_hectare(database, area_name, n_name, nha_name = "NHa")
```

Arguments

database	data.frame, data.table or any database
area_name	string with the name of field containing area in database
n_name	string with the name of field containing numbers of trees in database
nha_name	string with name you want for the field number of trees per hectare

generate_number_hectare_class

generate_number_hectare_class

Description

This function generates the NHa classes column which represents the classes with the number of surviving trees per hectare

Usage

```
generate_number_hectare_class(database, nha_field, n_field, area_field,
  plot_field, rounded_age_field, age_field, diameter_class_field, state_field,
  dap_field, amplitude = 2)
```

Arguments

database	data.frame, data.table or any database
nha_field,	name of the column containing the number of trees surviving per hectare
n_field,	name of the column containing the number of trees in a plot, or diameter class
area_field,	name of the column containing the area
plot_field,	name of the column containing the plots
rounded_age_field,	name of the column containing the rounded ages, example: 12,24,36

age_field,	name of the column containing the ages
diameter_class_field,	name of the column containing the diameter classes
state_field,	name of the column containing the states of trees
dap_field,	name of the column containing the diameter (DAP) of trees
amplitude,	integer integer number with the amplitude of diameter classes, example: 1, 2, 3

---

get_ages *get_ages*

Description

This function concatenates age values in a string for a query and returns the same

Usage

```
get_ages(database, age_name, age_values)
```

Arguments

database	data.frame, data.table or any database
age_name	string with the name of field (column) containing the ages
age_values	vector with the age values you want to assemble string to made query, example: c(12,24,36)

---

get_max *get_max*

Description

This function returns the maximum value of one or more fields of measurement variables

Usage

```
get_max(database, variables)
```

Arguments

database	data.frame, data.table or any database
variables	string vector with name(s) of the column (s) you want to know the maximum value

get_min — *get_min*

Description

This function returns the minimum value of one or more fields of measurement variables

Usage

```
get_min(database, variables)
```

Arguments

database	data.frame, data.table or any database
variables	string vector with name(s) of the column (s) you want to know the minimum value

get_place — *get_place*

Description

This function returns a database from a particular site or location present in the original database

Usage

```
get_place(database, place_name, place_value)
```

Arguments

database	data.frame, data.table or any database
place_name	string with the name of the column that represents the place
place_value	vector with values of that you want to filter the sites/places of the database

mount_list *mount_list*

Description

This function serves as a help for the check_diametric_increase function, it generates lists of values to be plotted as columns

Usage

```
mount_list(database, fieldColorName = "color", fieldQuantityName = "n",
  diameter_classes_name = "classedediametro")
```

Arguments

database,	data.table, data.frame or any database
fieldColorName,	string with the name you want for the color column
fieldQuantityName,	string name of column with number of trees
diameter_classes_name,	string name of column with the diameter classes

mount_query *mount_query*

Description

This auxiliary function checks that need to group fields of certain measurements

Usage

```
mount_query(database, select_names, group_names, option)
```

Arguments

database	data.frame, data.table or any database
select_names	string vector with the name(s) of the column(s) you want to include in the selection
group_names	string vector with the name(s) of the column(s) you want to group the results
option	options to make the query, can be 1,2,3 each one for one use in the analysis functions

plot_cols *plot_cols*

Description

This function serves as a help for the check_diametric_increase function, it generates a chart based on the lists created by the mount_list function

Usage

```
plot_cols(list, name, my_title, ceil = NaN)
```

Arguments

`list,`	list of columns generated by function mount_list with values and colors to be ploted
`name,`	string name with the name you want for the file png chart
`my_title,`	string with the name you want for the title of chart
`ceil,`	the maximum value to axis, this value will be applied in NHa or N

sort_columns_crescent *sort_columns_crescent*

Description

Sorts the database incrementally based on the selected column

Usage

```
sort_columns_crescent(database, column)
```

Arguments

`database`	data.frame, data.table or any database
`column`	string with the name of the column you want sort the database

Index

calculate_jumps_ages, 2
check_ages, 3
check_clones_different_plot, 3
check_dead_state, 4
check_diametric_increase, 4
check_existing_ages, 5
check_existing_place, 6
check_existing_plots, 6
check_measurement_ages, 7
check_measurements_state, 7
check_parcel_different_spacing, 8
check_size_age_plot, 8
check_undefined_spacing, 9
check_variables, 9
check_zero_measurement, 10
correct_probability, 10

find_missing_age, 11
find_missing_place, 11
find_missing_variable, 12

generate_colors_diameter_class, 12
generate_diameter_classes, 13
generate_initial_diameter_class, 13
generate_limiar, 14
generate_new_color, 14
generate_number_hectare, 15
generate_number_hectare_class, 15
get_ages, 16
get_max, 16
get_min, 17
get_place, 17

mount_list, 18
mount_query, 18

plot_cols, 19

sort_columns_crescent, 19

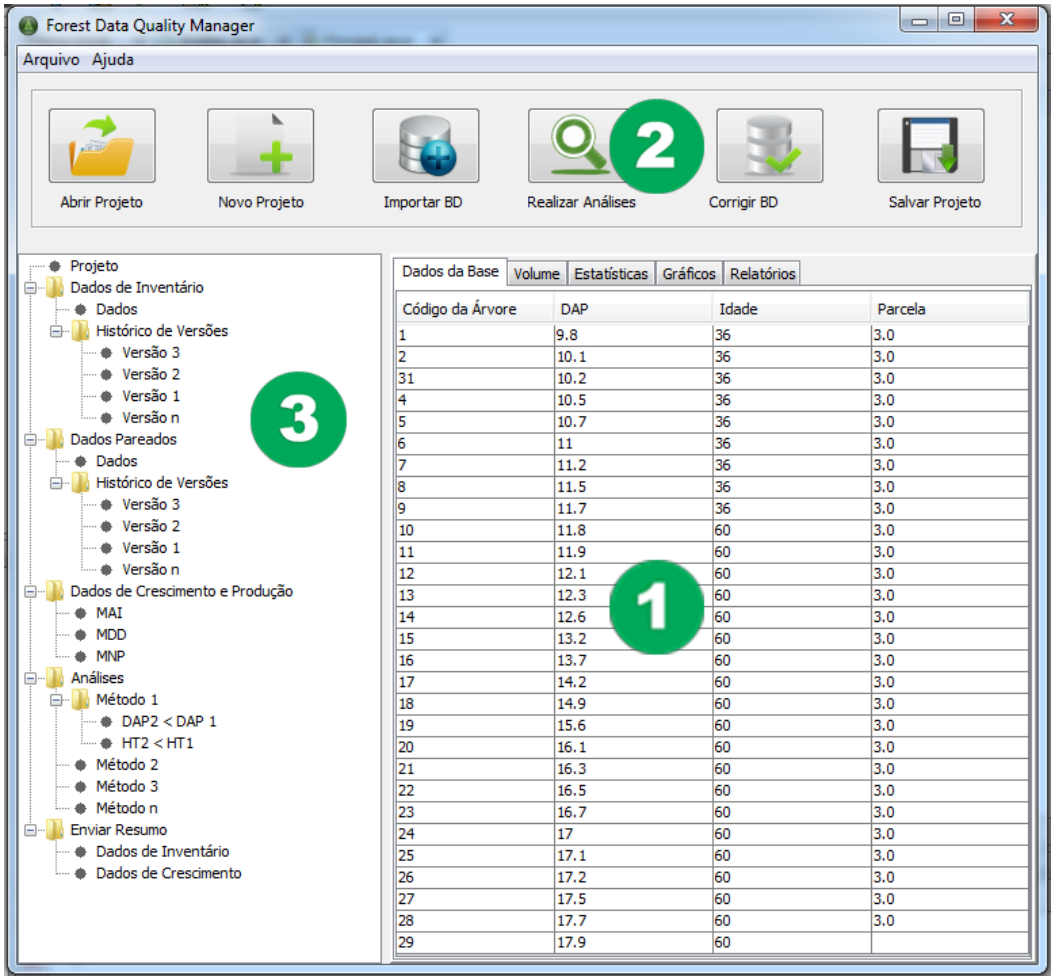
8.3 APÊNDICE C – Documentação do Software *Forest Data Quality Manager* (FDQM)

8.3.1 Implementação do Sistema

Finalizado o desenvolvimento e publicação do pacote FQD pretende-se implementar o sistema proposto, denominado *Forest Data Quality Manager* (FDQM). O sistema será desenvolvido utilizando a linguagem Java versão 8 (ORACLE, 2016).

O objetivo do sistema é proporcionar um ambiente que realize análises de bases de dados integrado ao pacote FDQ criado na etapa anterior, além disso, possibilitar a realização de correções existentes quanto a qualidade dos dados da base, correções que serão aplicadas sobre os erros obtidos como resultado das análises. Por fim, o sistema irá possibilitar uma gestão de conFiguração que cobrirá toda atividade de alteração da base de dados. Neste sentido, espera-se realizar o controle de versão de modo que cada alteração realizada e confirmada produza uma nova versão, mantendo o registro das versões anteriores e possibilitando ao usuário realizar a reversão caso deseje.

Vale destacar que a gestão de conFiguração proporciona também a gestão de mudanças, a fim de garantir que que as mudanças sejam aprovadas pelo usuário, de maneira integrada ao controle de versões. Tendo sido aprovada, uma base de dados poderá ser utilizada para seus devidos fins, inclusive para geração dos dados para modelos de crescimento e produção. A Figura 26 apresenta um protótipo da tela principal do sistema.

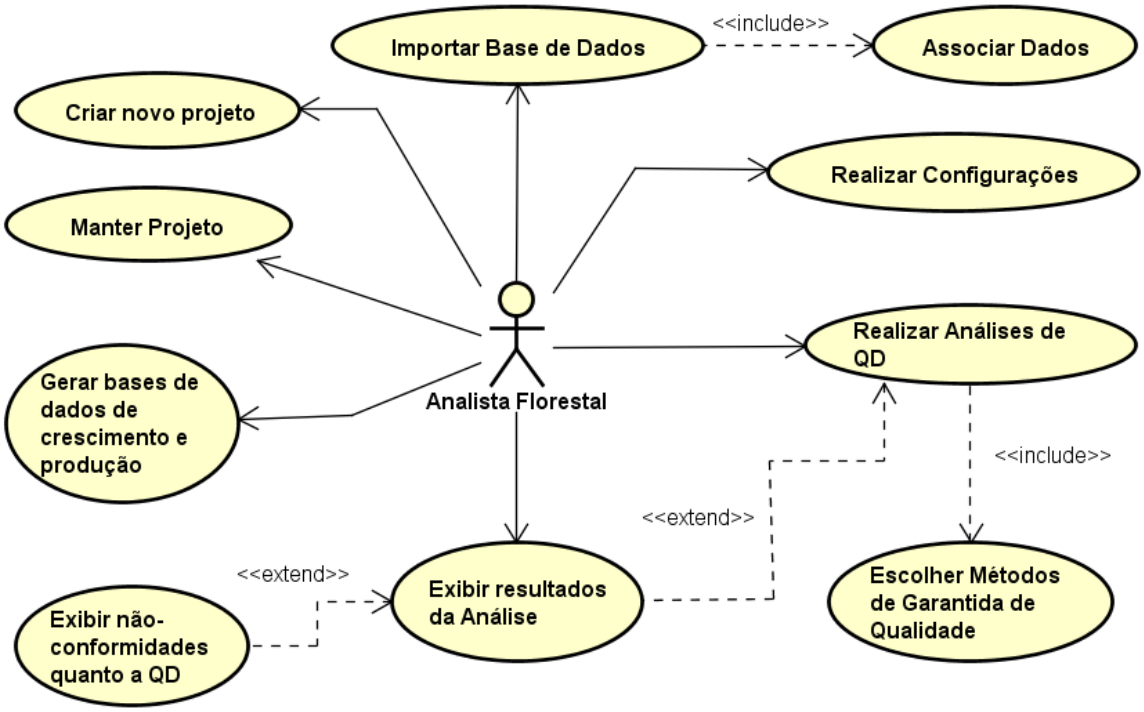


Na Figura acima apresentada é mostrada a interface principal do sistema, em (2) estão dispostas as funções que FDQM será possível de executar, as mesmas são descritas abaixo:

- Abrir projeto: Consiste na abertura de um projeto já criado anteriormente selecionando o arquivo desejado para trabalho.
- Novo projeto: Consiste na criação de um novo projeto, onde ele tem um nome e um local associado.
- Importar BD: Consiste na importação de novas bases de dados ou mesmo uma base inicial para trabalho, tal importação associa os dados da base aos dados que você deseja trabalhar dentro do projeto.
- Realizar Análises: Consiste na seleção dos métodos de verificação propostos na seção 3.7 deste trabalho, aqui o usuário poderá escolher as análises a serem realizadas sobre a sua base de dados de trabalho.

- Corrigir Base de Dados: Aqui o usuário irá escolher a base ou conjunto de dados que deseja efetuar correção, além disso o método de correção que será utilizado.
- Salvar projeto: Consiste em salvar o projeto atual do usuário.

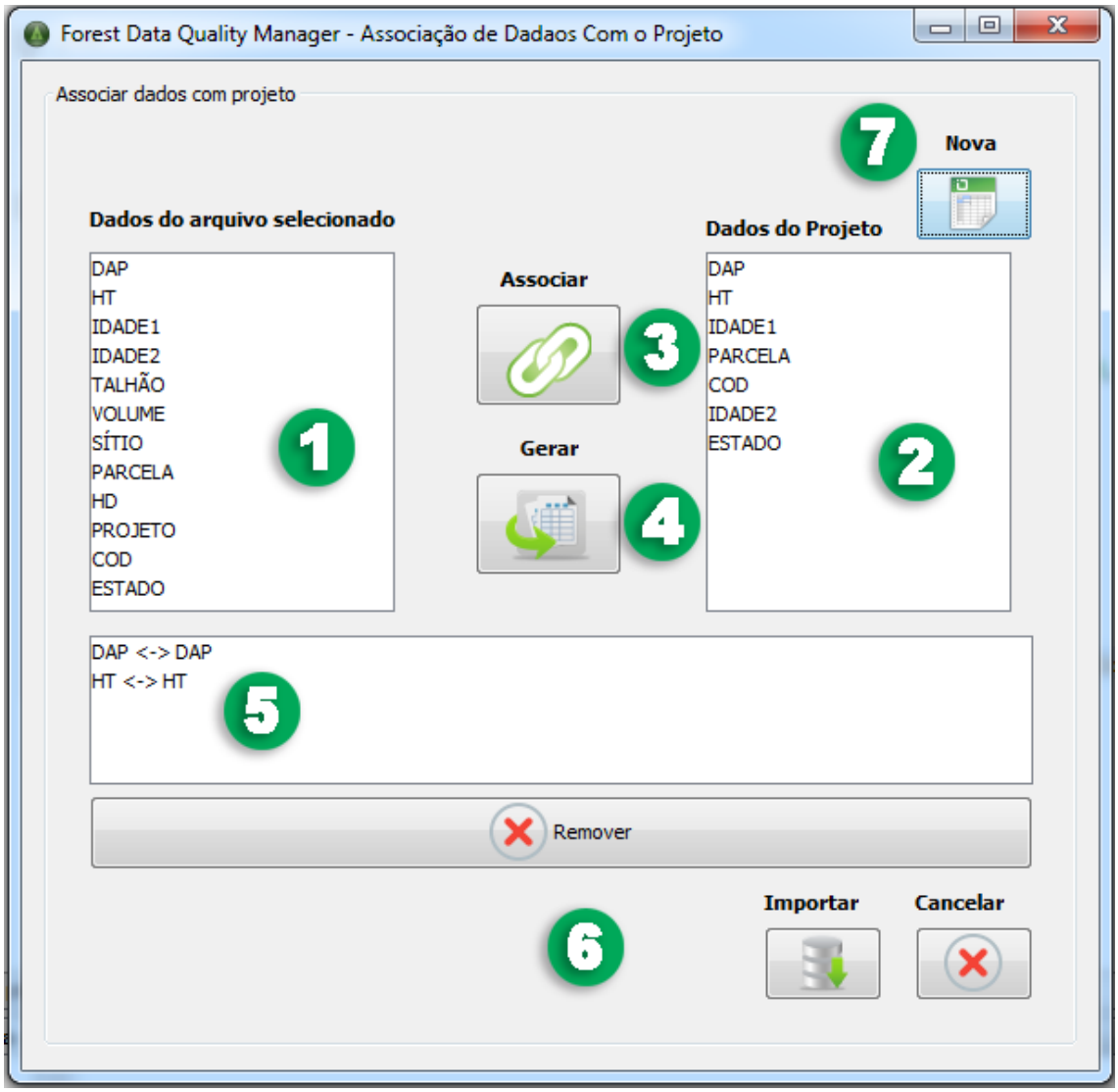
Em (1) é exemplificado como os dados serão mostrados para o usuário, através da tabela onde as colunas representam dados da base como código, DAP entre outros. Por fim em (3) se tem uma demonstração do projeto em forma de árvore, onde se têm um histórico e detalhamento completo sobre versões, dados, analises e correções realizadas. As Figuras abaixo dão uma visão geral do sistema por meio de diagramas de casos de uso, apresentando as funcionalidades que o usuário poderá utilizar.



A Figura acima apresenta a primeira parte das funcionalidades pertinentes ao FDQM, totalizando um total de 10 funcionalidades, que vão desde criar um novo projeto até exibir resultados de análise.

Começando pela importação das bases de dados, o analista florestal pode importar diferentes tipos de base de dados, podendo elas serem de um inventário florestal, ou de modelos de crescimento e produção (MDD, MNP, MAI). Após a realização, o usuário seleciona os dados a serem associados, ou seja, os dados da

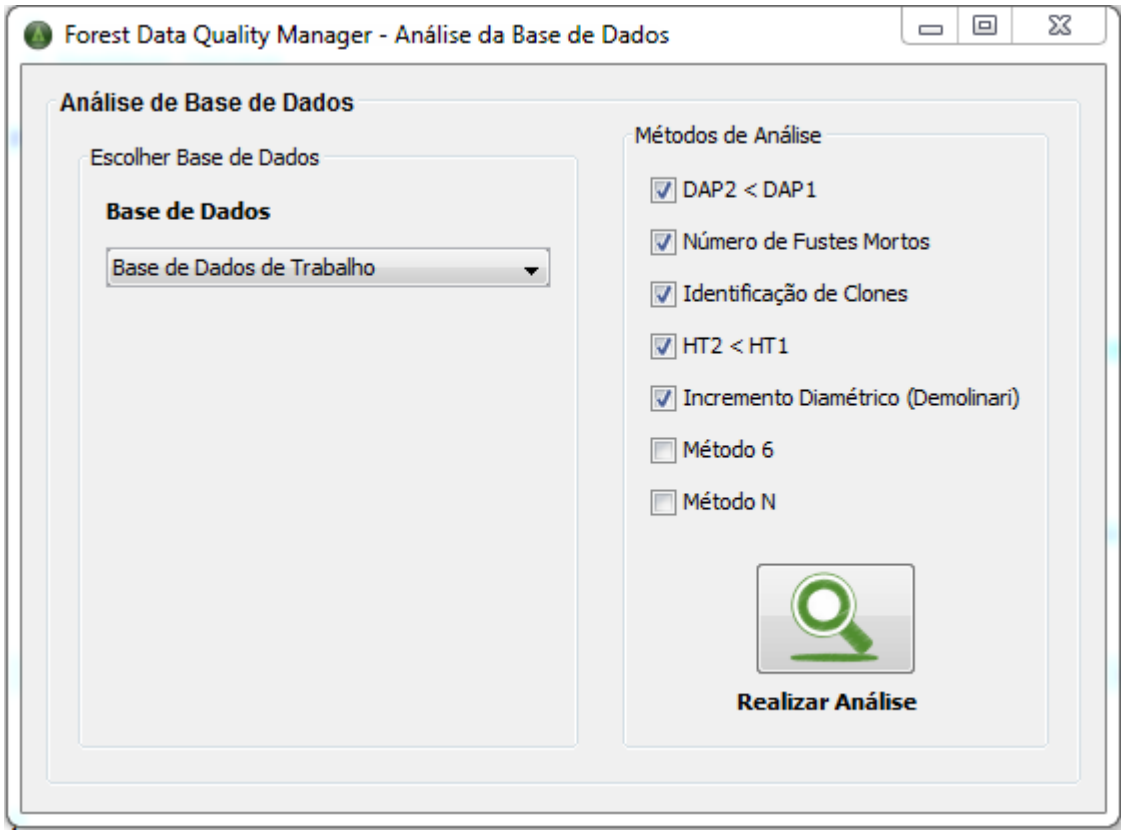
base que ele quer trabalhar, essas informações podem ser visualizadas na Figura abaixo.



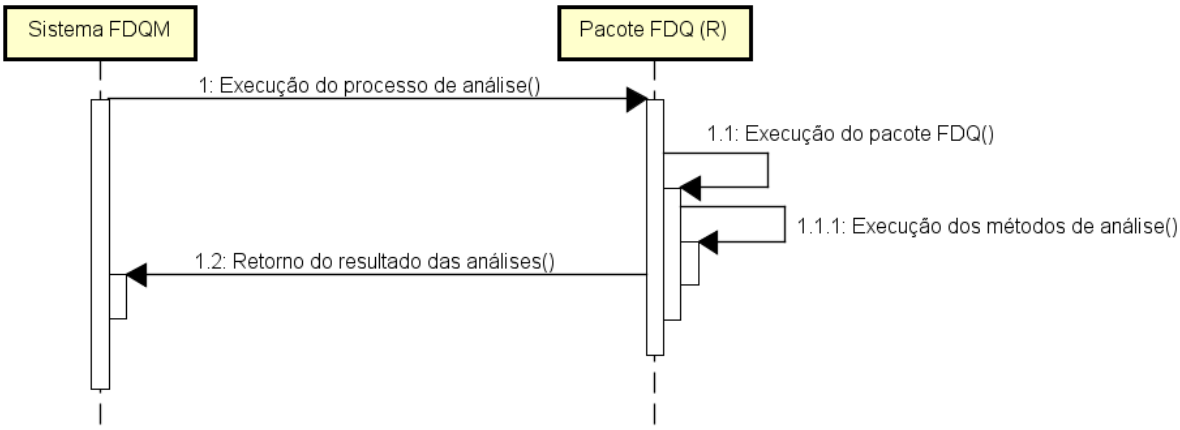
A Figura acima apresenta a associação de dados ao projeto, em 1 estão exemplificados os tipos de dados que o analista florestal pode escolher para trabalhar partindo de um arquivo externo. Em 2 são os dados já existentes no projeto. Em 3 ele cria a associação escolhendo quais variáveis ele deseja associar, por exemplo, em 5 ele associa o DAP e o HT do arquivo selecionado ao DAP e HT dos arquivos do projeto. Vale salientar que o usuário pode gerar uma variável que não está nem em arquivos nem no projeto através do como sinalizado em 4, em 6 ele poderá remover a associação que ele fez caso queira, por fim em 7 ele pode criar uma nova base fazendo um relacionamento automático entre as variáveis, isto acontece através dos

nomes das variáveis, se o nome da variável no projeto tiver o mesmo nome do arquivo então ele fará isso de forma automática.

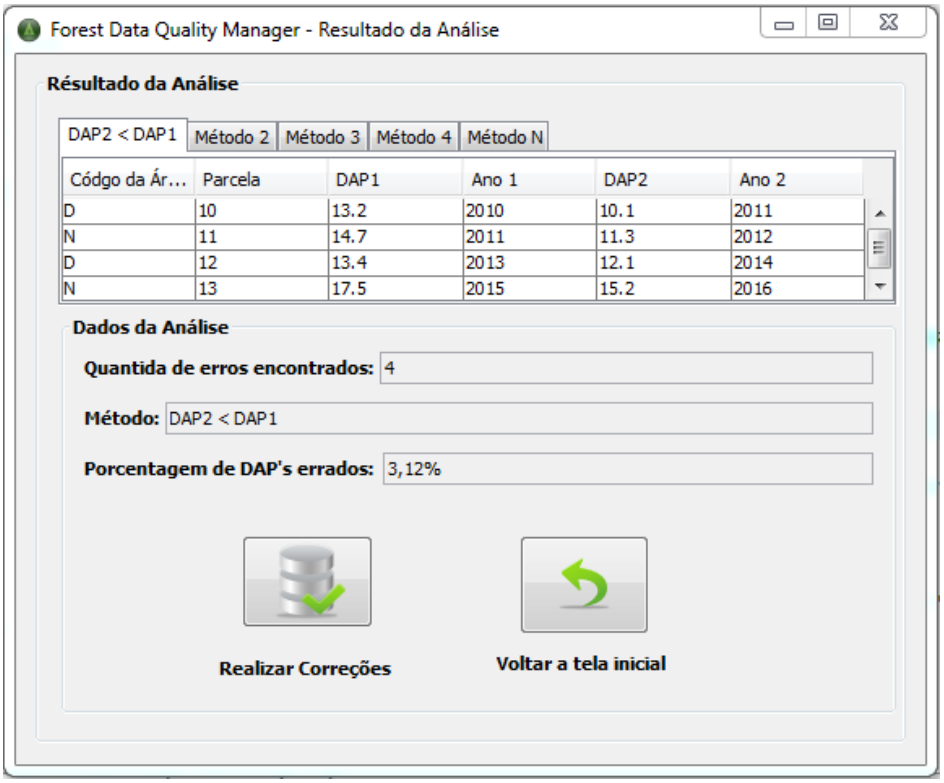
Finalizada a associação, pode-se realizar conFigurações, análises sobre a base de dados recém importada, porém para realização disso o usuário deverá escolher os métodos de garantia de qualidade que deseja que o sistema verifique na base. Por fim, após a realização da análise o usuário pode exibir os resultados da análise, nessa funcionalidade serão exibidas as não-conformidades encontradas em relação a análise realizada anteriormente. A Figura abaixo apresenta um protótipo da tela de análise do apresentando como o usuário irá fazer as análises.



A Figura acima apresenta o protótipo da tela de análises onde o usuário pode escolher a base de dados, seja ela, a de trabalho como selecionada ou outra base do projeto, ao lado temos a lista de métodos de análise a serem realizadas, tais como verificação de DAP1 < DAP2, entre outros. A Figura abaixo apresenta um diagrama de sequência de quando o usuário clicar em realizar a análise como será a comunicação do sistema com o pacote.

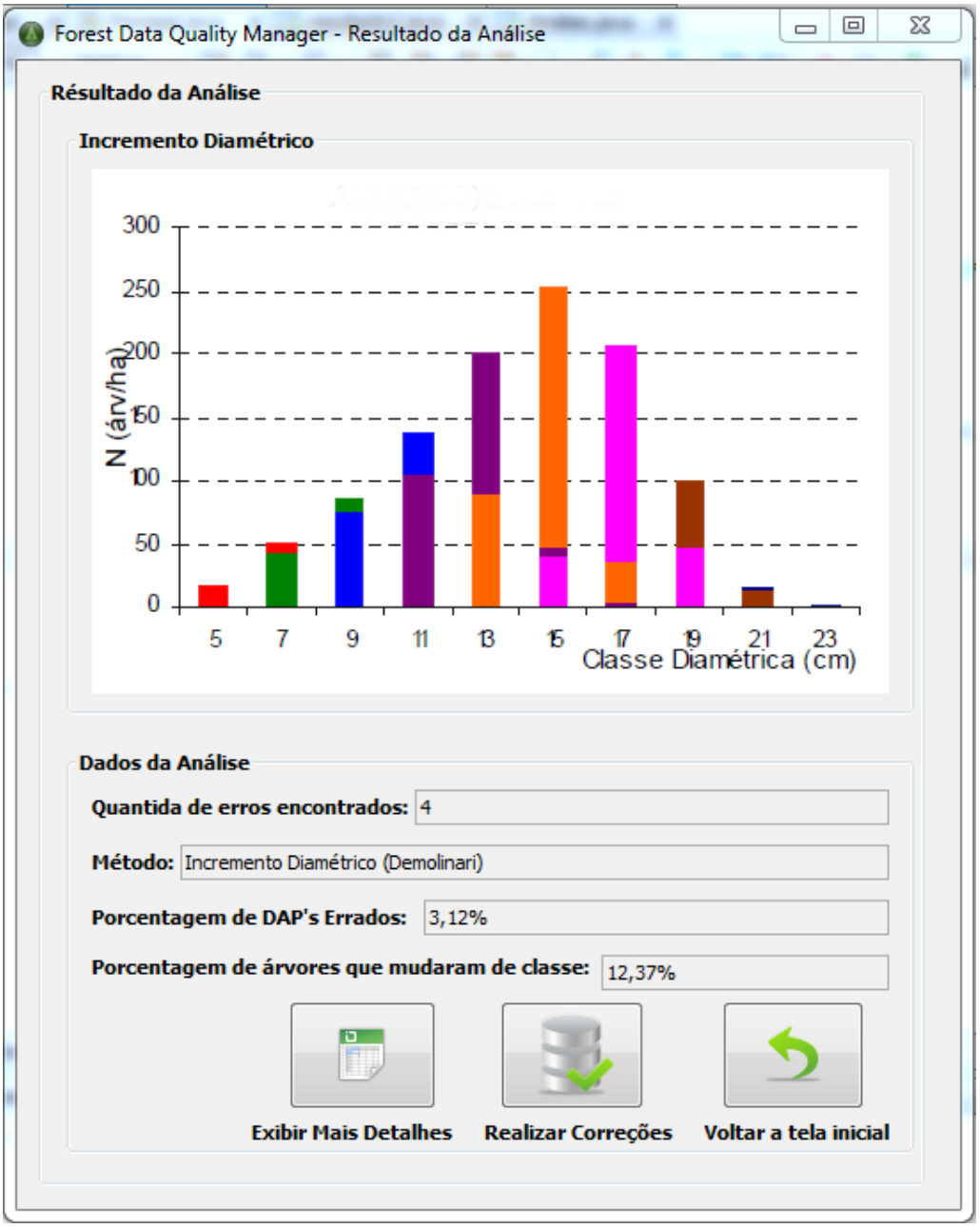


A Figura acima apresenta como ocorre o processo de análise entre o sistema (FDQM) e o pacote (FDQ), uma vez escolhidos os métodos de análise é feita a chamada de execução do R e então é executado o pacote e posteriormente a execução dos métodos escolhidos pelo usuário e então é feito o retorno das análises para o sistema para ser apresentado ao usuário. Uma vez finalizado o processo de análise é exibido o resultado da mesma ao usuário conforme as Figuras abaixo.



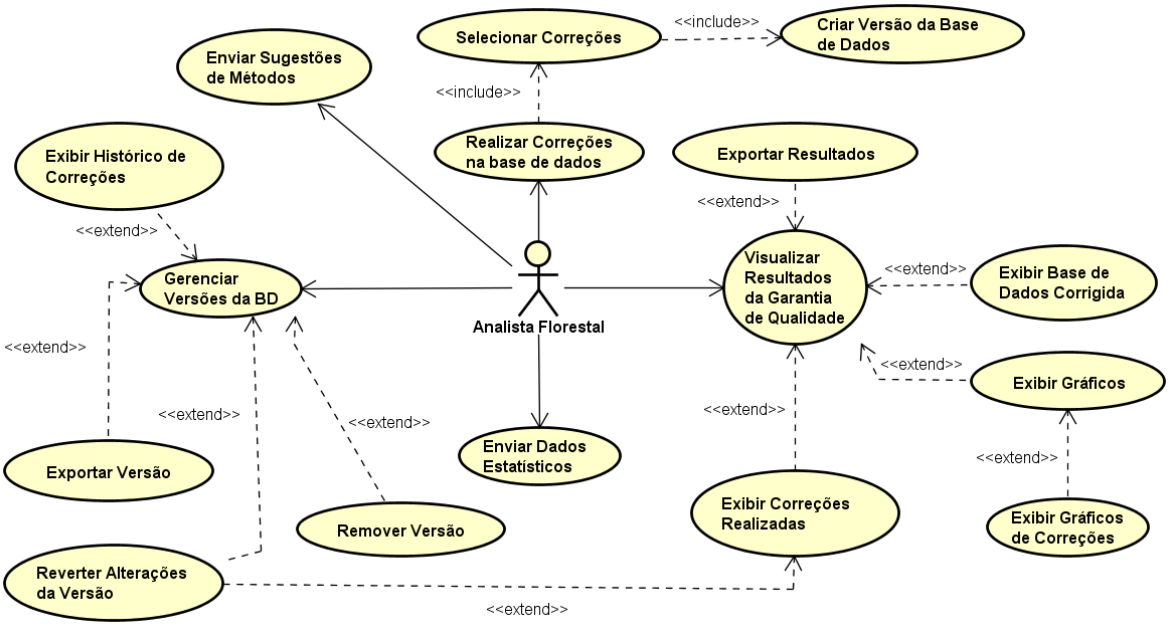
Na Figura acima pode-se observar o resultado da análise do método de verificação DAP2 < DAP1, contém a quantidade de erros encontrados desse tipo

além dos registros com erros exemplificados na tabela, além de apresentar a porcentagem de registros defeituosos do total.

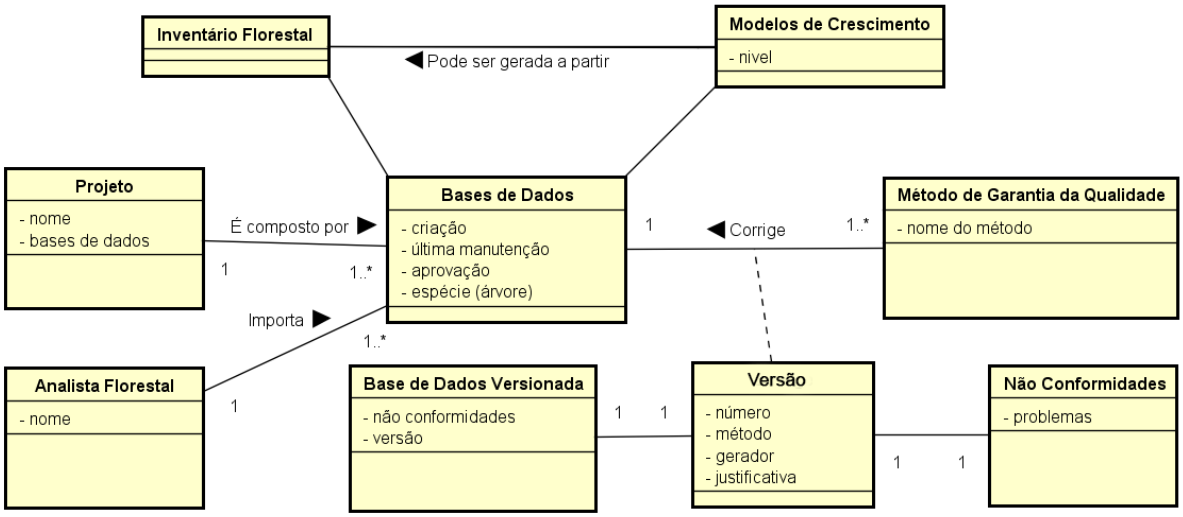


Já a Figura acima apresenta o resultado da análise pelo método do incremento diamétrico, demonstrando a quantidade de erros encontrados baseando-se nas árvores que mudaram de uma classe para outra diferindo drasticamente do que foi esperado, o gráfico é apresentado demonstra o comportamento das árvores em um período de tempo, e logo abaixo temos dados como quantidade de erros, porcentagem de DAP’s errados e de árvores que mudaram de classe. A Figura abaixo

apresenta o diagrama de caso de usos com o restante das funcionalidades pertinentes ao FDQM.



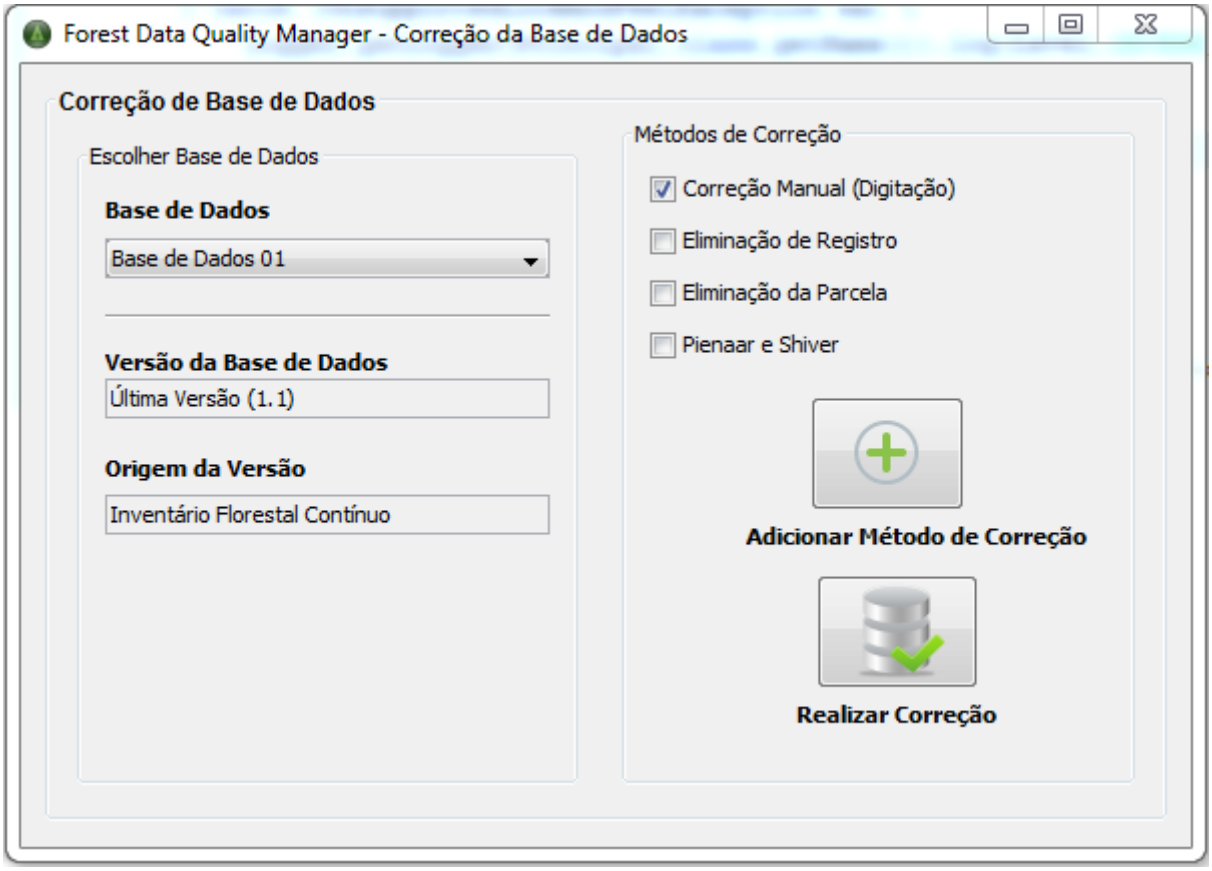
O diagrama de caso de uso da na Figura acima apresenta funções relacionadas após a análise da BD, iniciando pela correção, par realizar isso o usuário deverá selecionar quais correções ele deseja realizar. A Figura abaixo apresenta um modelo de domínio do sistema, apresentando as informações que o sistema vai gerenciar e seu domínio em si.



O modelo de domínio da Figura acima apresenta conceitos fundamentais do FDQM, incialmente temos o analista florestal, que é nosso principal usuário, ele

importa uma ou mais bases de dados para dentro do sistema, tais base (s) que compõe um projeto, vale ressaltar que as bases podem ser geradas por inventário florestal ou modelos de crescimento.

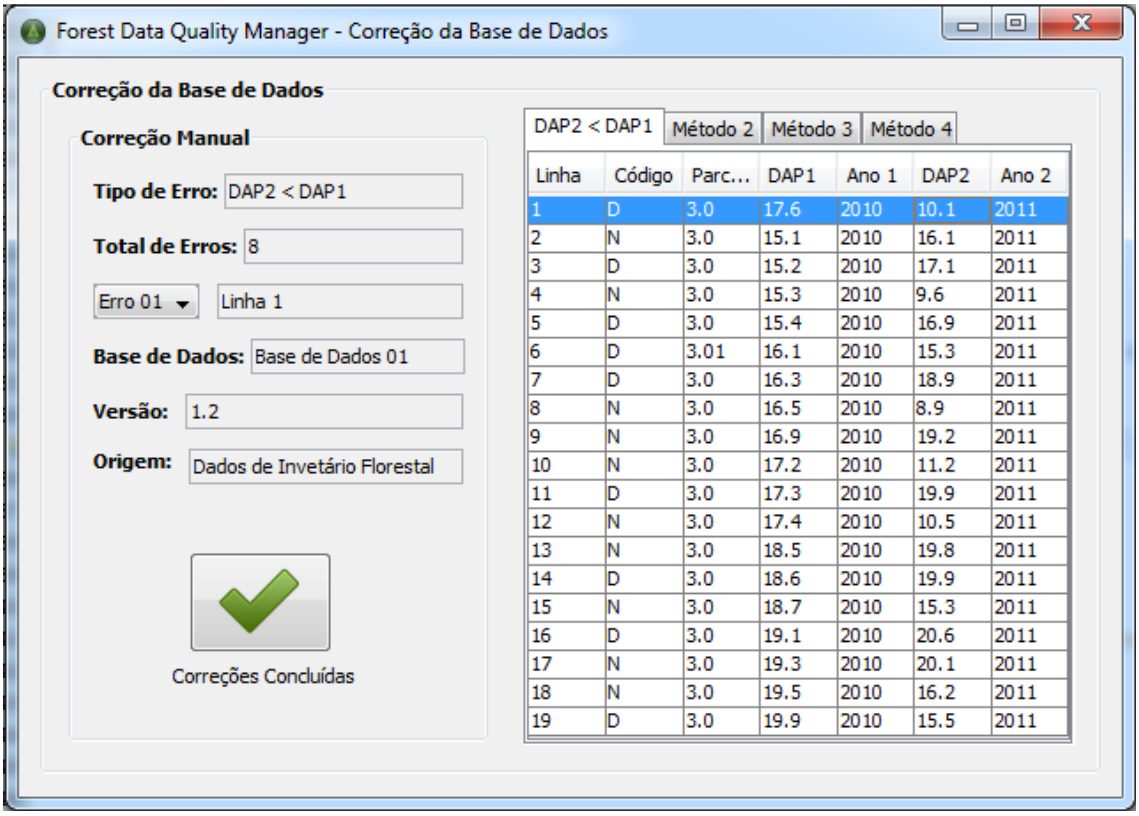
A ideia do sistema é proporcionar a essas N bases de dados métodos de garantia de qualidade, isso é representado no modelo de domínio através da classe associativa versão, que é o fruto do método de garantia de qualidade sobre uma base de dados, essa versão gera uma nova BD chamada de base de dados versionada, e além disso é gerado um relatório chamado de não-conformidades, o mesmo irá conter as não-conformidades (erros encontrados) em relação a qualidade de dados da BD. A Figura abaixo apresenta um protótipo da tela do processo de correção do FDQM.



É possível verificar na Figura acima a etapa de correção no sistema, onde o usuário terá acesso a última versão de sua base e além disso, escolherá os métodos de correção que ele quiser utilizar.

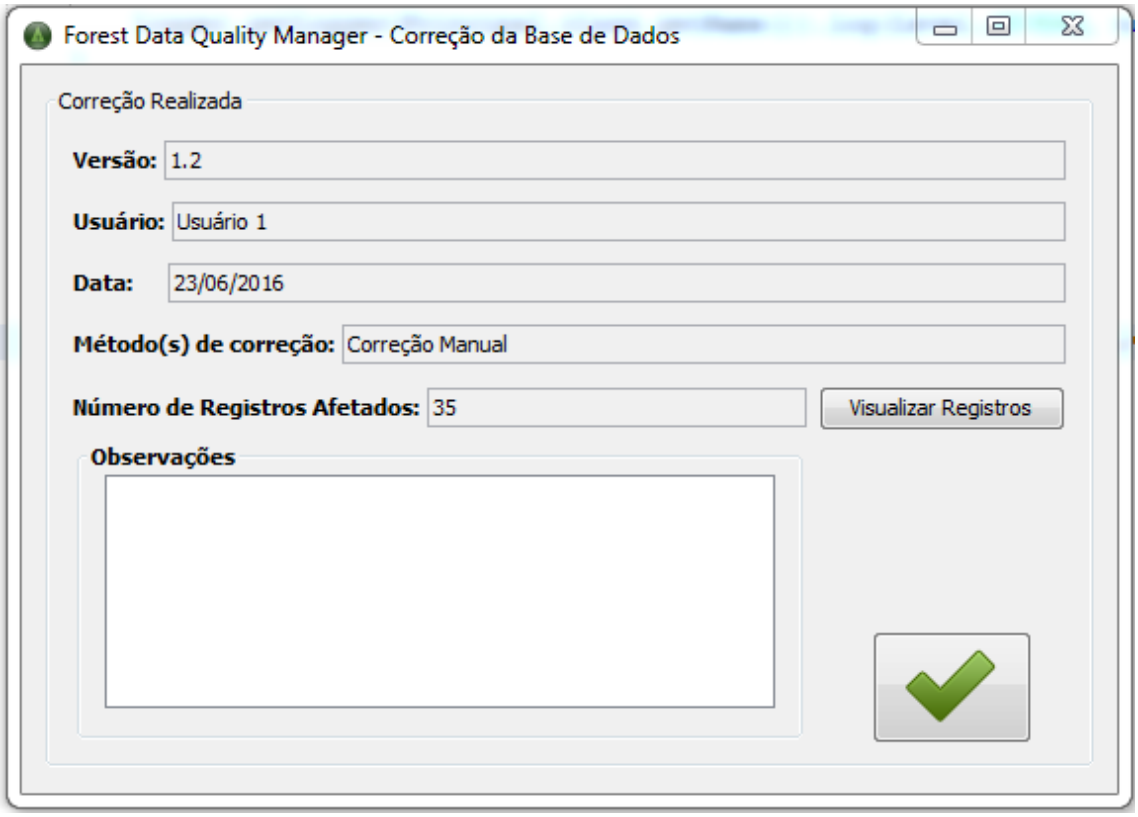
Os métodos acima foram proposto na seção 3.7 deste trabalho, a correção manual consiste no usuário digitar manualmente a correção desejada diretamente no campo do dado, as eliminações de registro e de parcela o usuário poderá selecionar os dados que deseja eliminar e confirmar a exclusão, o método de Pienaar e Shiver o usuário poderá utilizar para fazer estimativas de dados faltantes baseando-se nos existentes na mesma parcela referente ao dado que está faltando, e por fim o usuário através de plug-ins poderá adicionar novos métodos de correção ao sistema.

Ressaltando que as correções partem dos erros encontrados na etapa de análise da BD, após realizada a correção é gerada automaticamente uma nova versão da base, onde o versionamento entra em ação é importante destacar que após cada correção realizada uma nova versão é gerada pelo sistema. Ao fim das correções o usuário pode visualizar o escopo das correções, seja através da exibição da BD corrigida ou através de gráficos. Uma vez escolhido o método de correção a Figura abaixo exemplifica como funcionará a correção manualmente.

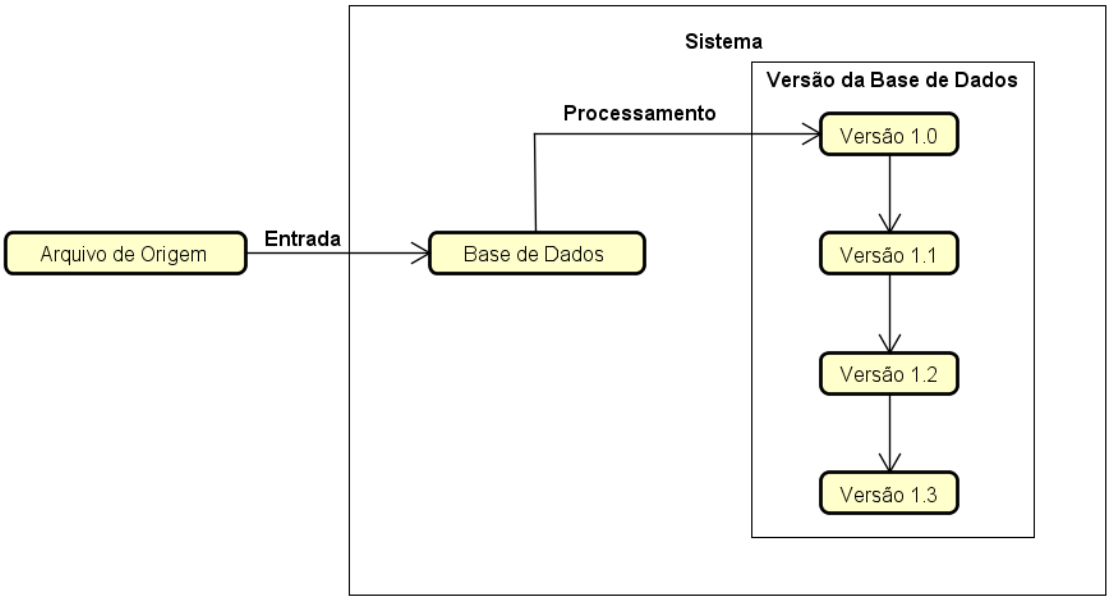


Linha	Código	Parc...	DAP1	Ano 1	DAP2	Ano 2
1	D	3.0	17.6	2010	10.1	2011
2	N	3.0	15.1	2010	16.1	2011
3	D	3.0	15.2	2010	17.1	2011
4	N	3.0	15.3	2010	9.6	2011
5	D	3.0	15.4	2010	16.9	2011
6	D	3.01	16.1	2010	15.3	2011
7	D	3.0	16.3	2010	18.9	2011
8	N	3.0	16.5	2010	8.9	2011
9	N	3.0	16.9	2010	19.2	2011
10	N	3.0	17.2	2010	11.2	2011
11	D	3.0	17.3	2010	19.9	2011
12	N	3.0	17.4	2010	10.5	2011
13	N	3.0	18.5	2010	19.8	2011
14	D	3.0	18.6	2010	19.9	2011
15	N	3.0	18.7	2010	15.3	2011
16	D	3.0	19.1	2010	20.6	2011
17	N	3.0	19.3	2010	20.1	2011
18	N	3.0	19.5	2010	16.2	2011
19	D	3.0	19.9	2010	15.5	2011

A Figura acima mostra o resultado das análises e a quantidade de erros que a mesma possui, indicando a linha onde se encontra o respectivo (linha em azul), uma vez que o erro é mostrado ao usuário o mesmo pode trocar digitar diretamente na tabela o dado correto e por fim concluir a correção. Pode-se observar na Figura abaixo o processo de finalização da correção.



Apresenta-se na Figura acima mostra como será o processo de documentar a correção, exibindo uma tela que contém informações como versão, usuário que realizou, data, o método (s) de correções, e o usuário ainda pode colocar observações sobre o que foi corrigido entre outras. Após a realização de todo processo vale destacar que o sistema proporciona ao usuário uma “linha do tempo” da base de dados, apresentando todas as versões geradas e detalhadamente cada mudança realizada contendo o que mudou, quem mudou, data de modificação, entre outras características da gestão de mudança. A Figura abaixo apresenta como versionamento funciona dentro do DQFM.



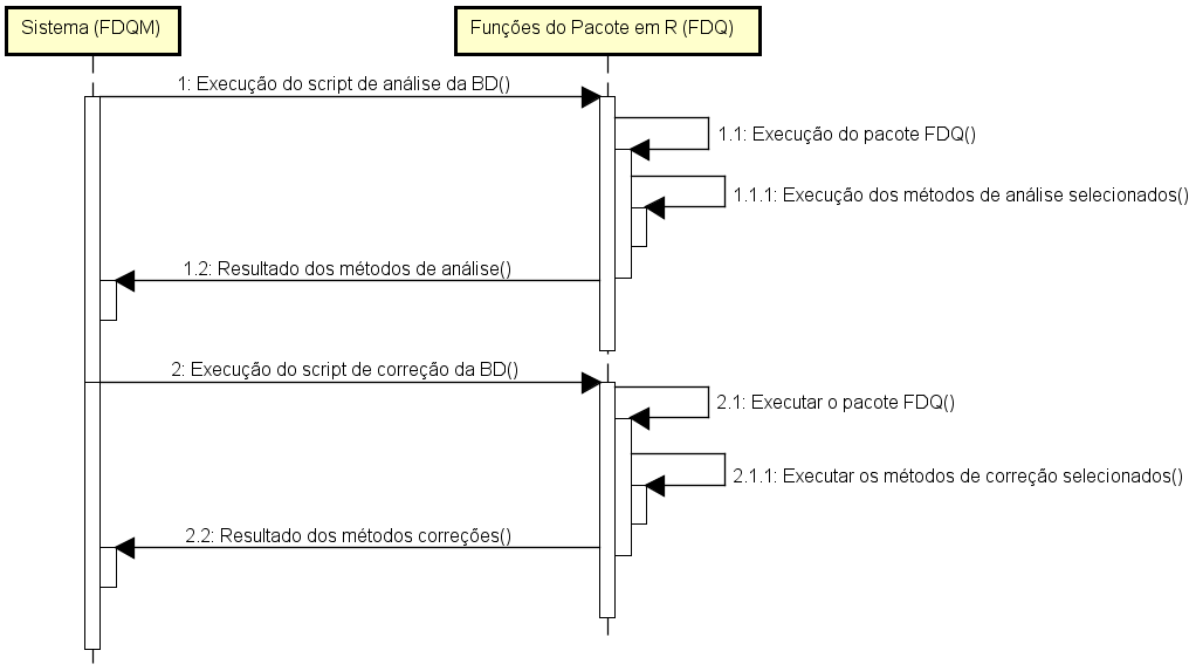
A Figura acima apresenta em um diagrama de atividades como versionamento irá funcionar no sistema, o arquivo de origem provindo do usuário é importado para o sistema e passa por um processamento (métodos de garantia de qualidade), a partir disso são geradas versões das modificações realizadas, vale ressaltar que cada mudança realizada na base de dados gera uma nova versão da mesma.

Por fim, por meio da gestão de conFiguração implementada, e o detalhamento da base que ela proporciona, o sistema disponibiliza ao analista florestal a possibilidade pode reverter alterações realizadas em alguma correção realizada, isso ocorre pelo mecanismo de *rollback* que atua também como medida de segurança da BD.

8.3.2 Integração do FDQ com o sistema FDQM

Finalizada a implementação do sistema o próximo passo é realizar a integração do pacote criado (FDQ) utilizando a linguagem R realizado na primeira etapa com o sistema FDQM. A fim de solucionar isso utilizaremos um o plug-in rConector, desenvolvido para o software Forest Measure (FILHO et al., 2014). De acordo com Romualdo (2015) com o uso do plug-in é possível realizar a execução dos scripts e funções no ambiente R, buscar informações sobre os dados em tempo de execução, buscar resultados obtidos, ou seja, manipular o software R(R DEVELOPMENT CORE

TEAM, 2005) a partir do sistema proposto. A Figura abaixo apresenta o funcionamento da conexão do sistema FDQM com o pacote FDQ criado no R.



A Figura acima apresenta a execução dos processos entre o sistema (FDQM) e o pacote (FDQ), em (1) temos a execução dos scripts de análise da BD, que contém os métodos de garantia da qualidade de dados escolhidos pelo usuário, já com a requisição do sistema no R temos a execução do pacote e posteriormente dos métodos selecionados do pacote, após isso é retornada a análise para o sistema (1.2). Após a rotina de execução de todas as análises em (2) temos o início dos scripts de correção da BD, no R é executado novamente o pacote (FDQ) e posteriormente os métodos de correção selecionados. Por fim em (2.2) temos o retorno dos resultados obtidos da aplicação dos métodos de correção.

Contudo, além de integrar o FDQ ao sistema para usufruir das funções contidas nele, temos a necessidade de integrar pacote o Fgmutils (FILHO et al., 2015) que tem como objetivo fornecer funções para manejo de bases de dados florestais como geração de base pareada, o pacote trabalha diretamente com variáveis características como DAP, HT, entre outras, permitindo a manipulações de dados presentes na base.