Modelos Heterogêneos para a Previsão de Safras e Qualidades de Cultivo na Indústria Sucroenergética

Geraldo Gomes da Cruz Júnior

Universidade Federal Rural de Pernambuco
Departamento de Estatística e Informática
Programa de Pós-Graduação em Informática Aplicada

Recife
2019

Geraldo Gomes da Cruz Júnior

Modelos Heterogêneos para a Previsão de Safras e Qualidades de Cultivo na Indústria Sucroenergética

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Informática Aplicada da Universidade Federal Rural de Pernambuco como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Prof. Dr. Gabriel Alves de Albuquerque Junior
Coorientador: Prof. Dr. Jones Oliveira de Albuquerque

Recife
2019

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Central, Recife-PE, Brasil

C957m Cruz Júnior, Geraldo Gomes da
Modelos heterogêneos para a previsão de safras e qualidades de cultivo na indústria sucroenergética / Geraldo Gomes da Cruz Júnior. – 2019.
151 f.: il.

Orientador: Gabriel Alves de Albuquerque Junior.
Coorientador: Jones Oliveira de Albuquerque.
Dissertação (Mestrado) – Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, BR-PE, 2019.
Inclui referências e anexo(s).

1. Processo estocástico 2. Markov, Processos de 3. Monte Carlo, Método de 4. Processo decisório - Modelos matemáticos 5. Indústria açucareira 6. Agricultura e energia I. Albuquerque Junior, Gabriel Alves de, orient. II. Teixeira, Jones Oliveira de, coorient. III. Título

CDD 004

GERALDO GOMES DA CRUZ JÚNIOR

MODELOS HETEROGÊNEOS PARA A PREVISÃO DE SAFRAS E QUALIDADES DE CULTIVO NA INDÚSTRIA SUCROENERGÉTICA

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Informática Aplicada do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco, como requisito para obtenção do Grau de Mestre em Informática Aplicada.

Aprovada em: 20 de Fevereiro de 2019.

BANCA EXAMINADORA

Gabriel Alves de Albuquerque Junior (Orientador)
Universidade Federal Rural de Pernambuco - UFRPE
Departamento de Estatística e Informática – DEINFO

Gustavo Rau de Almeida Callou
Universidade Federal Rural de Pernambuco - UFRPE
Departamento de Estatística e Informática – DEINFO

Rafael Melo Macieira
Serviço Nacional de Aprendizagem Industrial – SENAI
Instituto SENAI de Inovação para Tecnologias da Informação e Comunicação - ISITICs

A Deus,
à minha família,
aos meus professores,
à equipe do ISI-TICs,
e a todos os meus amigos.

Agradecimentos

Agradeço primeiramente a Deus por ter me guiado, protegido e iluminado durante toda a minha vida e ter me possibilitado todas as condições para a conclusão deste trabalho.

Agradeço muito a toda a minha família e em especial aos meus pais, Geraldo Gomes e Nadja Carneiro, que com muito amor sempre me incentivaram e me fizeram reconhecer o valor do estudo como uma ferramenta transformadora do mundo e realizadora de sonhos.

Agradeço da forma mais especial possível a minha vó Rita (*in memorian*), a melhor pessoa do mundo, que tanto torceu por mim, mas que infelizmente se foi antes de ver o término desta etapa.

Agradeço à Universidade Federal Rural de Pernambuco e a todos os professores do PPGIA e do curso de BSI, por serem os meus maiores incentivadores na busca pelo conhecimento e aprendizado que tive durante a vida acadêmica. Em especial agradeço ao meu orientador, professor Gabriel Alves, e ao meu coorientador, professor Jones Albuquerque, pelo voto de confiança, por todo incentivo, cobrança, orientação e paciência durante o desenvolvimento deste trabalho. Um agradecimento especial também para Eduardo Chaves, um amigo sempre solícito, até durante as férias, para resolver qualquer situação administrativa do programa.

Agradeço à toda a equipe do ISI-TICs, por todo o aprendizado e companheirismo. O Instituto possibilitou uma imersão no universo de inovações industriais, mostrando como ferramentas de TICs são fundamentais para mudanças de impacto nacional e mundial. Um agradecimento especial à melhor sala!

Agradeço muito a todos os meus amigos, novos ou velhos, da escola, ETEPAM, graduação, mestrado e vida que são peças essenciais para eu ser o que sou, em especial à André Luiz, Maykel Berenguer, Débora Freitas, Harmando Coutinho, Mário Gomes, Carlos César, Larissa Melo e Rafaella Leandra. Agradeço também por toda paciência e ajuda da minha namorada e melhor amiga Elyenay Bandeira.

Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), à equipe da São José Agroindustrial, ao Departamento de Agronomia da UFRPE e a todo o meu referencial teórico. Obrigado por fazerem parte da minha vida!

"Seja fiel as pequenas coisas porque é nelas que mora a sua força."
(Madre Teresa)
"O estudo da complexidade vai ser a ciência do século XXI"
(Stephen Hawking)

Resumo

A agricultura é um importante setor para a economia e sociedade brasileira e mundial. As agroindústrias junto com pequenos agricultores, são responsáveis por colocar o país entre os maiores produtores deste setor no mundo. Um importante representante da agroindústria brasileira são as usinas sucroenergéticas, responsáveis pelo plantio, cultivo e produção de derivados da cana-de-açúcar. A cana-de-açúcar é responsável por 80% de todo o açúcar produzido, sendo o Brasil o maior produtor de cana no mundo. A agricultura vem ganhando muita força, produtividade e inovação, principalmente quando alinhada a conceitos de tecnologias recentes. A agricultura de precisão faz uso de recursos inovadores de softwares e hardwares para avaliar e monitorar as condições de cultivos. Com isso, gera um grande conjunto de dados preparados para serem analisados e posteriormente servirem para a tomada de decisão. Porém, as agroindústrias normalmente utilizam grandes áreas para o plantio de diferentes tipos de monoculturas. A dimensão dessas áreas geralmente acarreta grandes problemas para o monitoramento, controle de pragas, meio ambiente e cultivo da plantação. Visando maximizar a produção enquanto se procura minimizar os impactos ambientais, o agronegócio procura cada vez mais investir em tecnologias e pesquisas que auxiliem na análise de diversos dados coletados ao longo das safras. Esses dados precisam ser tratados para se extrair informações relevantes. Tendo em vista essa necessidade, este trabalho utiliza o CRISP-DM como processo para a mineração dos dados, o qual é uma ferramenta muito útil como parte da definição da metodologia do projeto. Com os levantamentos realizados e revisão da literatura, percebe-se que é necessário fazer previsões e análises de comportamentos futuros a partir de objetos modelados. Contudo, a qualidade das respostas de um modelo qualquer proposto depende da precisão da estrutura computacional e dos dados que alimentam o modelo. A problemática levantada se trata de um sistema complexo, ou seja, um sistema que é composto por inúmeros elementos que interagem, de modo que o comportamento agregado não pode ser inferido do comportamento das unidades constituintes isoladamente, logo se enquadrando na dinâmica de cultivo de um canavial, abrangendo diferentes variáveis de interesse, como temperatura, luminosidade, solo, dentre outras. Para a compreensão

da problemática e desenvolvimento das propostas de solução utilizou-se a metodologia do Tookit HCD, um processo com princípios focados na imersão para compreensão da problemáticas e desenvolvimento de soluções alinhadas com as necessidades dos usuários finais. Esta dissertação discute e sugere um modelo piloto para implantação de uma infraestrutura para coleta, armazenamento, processamento e visualização de dados das plantações através da utilização da internet das coisas e de dispositivos móveis. A contribuição parte da apresentação de dois modelos estocásticos para a comunidade, focados na predição de safras, índices de qualidade e cenários de cultivo das diferentes fases de crescimento da cana-de-açúcar. Os modelos obtiveram resultados positivos e promissores nas simulações realizadas. Utilizou-se dados da Usina Agroindustrial São José, em uma área de estudo de 15 hectares. O primeiro modelo apresentado, baseado na utilização do método de Monte Carlo em cadeias de Markov, obteve bons resultados na predição de safras e índices de qualidade do canavial. O modelo obteve em experimentos testes de hipótese com *p-values* de 0.8754 e coeficientes kappa de 0.68. O Outro modelo é embasado em autômatos celulares estocástico, o qual visa a simulação de cenários georreferenciados da plantação, também classificando regiões como boas, más ou medianas. O Modelo conseguiu em experimentos um *p-value* de 0.8635 no teste de hipótese e um coeficiente kappa de 0.71. Em ambos os modelos os *p-values* e kappas indicam uma relação positiva entre os resultados dos modelos e os dados da base. O investimento em tecnologias e inovações agrícolas é essencial para otimizar as produções desse setor, bem como reduzir gastos enquanto se preserva o meio ambiente.

Palavras-chave: Autômatos Celulares Estocásticos. Cadeias de Markov. Método de Monte Carlo. Canaviais. Cana-de-açúcar. Plantações em nuvem.

Abstract

Agriculture is an important sector for the Brazilian and world economy and society. The agroindustries together with small farmers are responsible for placing the country among the largest producers of this sector in the world. An important representative of the Brazilian agro-industry are the sugar-energy plants, responsible for the planting, cultivation and production of sugarcane by-products. Sugarcane accounts for 80 % of all sugar produced, Brazil being the largest producer of sugarcane in the world. Agriculture has been gaining a lot of strength, productivity and innovation, especially when aligned with recent technology concepts. Precision farming makes use of innovative software and hardware resources to evaluate and monitor crop conditions. With this, it generates a large set of data prepared for analysis and later serve for decision making. However, agroindustries usually use large areas for the planting of different types of monocultures. The size of these areas usually poses major problems for monitoring, pest control, the environment and planting. In order to maximize production while seeking to minimize environmental impacts, agribusiness increasingly seeks to invest in technologies and research that assist in the analysis of various data collected throughout the harvests. These data need to be processed to extract relevant information. Given this need, this work uses CRISP-DM as a process for data mining, which is a very useful tool as part of the definition of the project methodology. With the realized surveys and literature review, it is realized that it is necessary to make predictions and analyzes of future behaviors from modeled objects. However, the quality of the answers of any proposed model depends on the precision of the computational structure and the data that feed the model. The problematic raised is a complex system, that is, a system that is composed of numerous elements that interact, so that the aggregate behavior can not be inferred from the behavior of the constituent units alone, thus falling within the dynamics of cultivating a covering different variables of interest, such as temperature, luminosity, soil, among others. In order to understand the problematic and development of the solution proposals, the methodology of Tookit HCD was used, a process with principles focused on immersion to understand the problems and development of solutions in line with the needs of the end users. This dissertation

discusses and suggests a pilot model for implementing an infrastructure for collection, storage, processing and visualization of plantation data through the use of the internet of things and mobile devices. The contribution is based on the presentation of two stochastic models for the community, focused on the prediction of harvests, quality indexes and cultivation scenarios of the different stages of sugarcane growth. The models obtained positive and promising results in the simulations. Data from the São José Agroindustrial Plant was used in a study area of 15 hectares. The first model presented, based on the use of the Monte Carlo method in Markov chains, obtained good results in the prediction of crops and quality indices of sugar cane. The model obtained in experiments hypothesis tests with p-values of 0.8754 and kappa coefficients of 0.68. The other model is based on stochastic cellular automata, which aims to simulate georeferenced scenarios of the plantation, also classifying regions as good, bad or medium. The model obtained in experiments a p-value of 0.8635 in the hypothesis test and a kappa coefficient of 0.71. In both models the p-values and kappas indicate a positive relationship between the model results and the base data. Investment in agricultural technologies and innovations is essential to optimize production in this sector, as well as reduce costs while preserving the environment.

Keywords: Stochastic Cellular Automata. Markov chains. Monte Carlo method. Canebrake. Sugar cane. Cloud Plantations.

Lista de ilustrações

Lista de tabelas

Lista de algoritmos

Lista de siglas

AC Autômatos Celulares

ACE Autômatos Celulares Estocásticos

ACP Autômatos Celulares Probabilísticos

AM Aprendizado de Máquina

ATR Açúcar Teórico Recuperável

CRISP-DM *CRoss Industry Standard Process for Data Mining*

DE Desafio Estratégico

EUPS Equação Universal de Perda do Solo

FAO *Food and Agriculture Organization*

HCD *Human-Centered Design*

IA Inteligência Artificial

IBGE Instituto Brasileiro de Geografia e Estatística

INPE Instituto Nacional de Pesquisas Espaciais

IQR *Interquartile Range*

ISI Instituto SENAI de Inovação

IoT *Internet of Things*

KDD *Knowledge Discovery in Databases*

LOB *Lower Outlier Boundary*

MCE *Multi-Criteria Evaluation*

MCH Monte Carlo Híbrido

MCMC Cadeia de Markov de Monte Carlo

MD Mineração de Dados

MMC Método de Monte Carlo

MSC Morte Súbita dos Citros

PCC Pol da Cana Corrigido

PIB Produto Interno Bruto

RIF Risco de Incêndios Florestais

SENAI Serviço Nacional de Aprendizagem Industrial

SIG Sistemas de Informações Geográficas

TCH Toneladas de Cana por Hectare

TMPH Toneladas de Muda por Hectare

TPH Toneladas de Pol (açúcar) por Hectare

UOB *Upper Outlier Boundary*

Sumário

Capítulo 1

Introdução

A agroindústria pode ser compreendida como o mercado que relaciona o comércio e a indústria da cadeia produtiva agrícola ou pecuária (MATOS; PESSOA, 2011). No Brasil, o setor agrícola tem grande importância para a economia e sociedade, sendo este responsável por grande parte do Produto Interno Bruto (PIB) nacional, colocando o Brasil entre os maiores produtores do mundo neste setor (COSTA; GUILHOTO; IMORI, 2013) (GUILHOTO et al., 2011).

Os grandes sistemas de plantio brasileiros fazem uso da monocultura para as suas produções. A monocultura é uma técnica que substitui a cobertura vegetal original de uma área por uma única cultura, uma prática que já é realizada no Brasil desde o século XVI (SILVA; FERREIRA, 2018).

A monocultura, apesar de ser utilizada para otimizar a produção, traz uma série de possíveis impactos negativos para o solo e para o meio ambiente. O desmatamento, a queimada e o cultivo em uma grande área prejudica o solo, retirando nutrientes e o deixando pobre. Além disso, monoculturas são mais sujeitas a pragas e ao consumo de recursos naturais excessivos, como o de água e o de energia elétrica. Estes pontos geram um grande problema para as empresas, população e principalmente para o meio ambiente (LIMA et al., 2017).

Um importante representante da agroindústria brasileira são as usinas sucroenergéticas, responsáveis pelo plantio, cultivo e produção de derivados da cana-de-açúcar. O Brasil é o principal produtor da cana no mundo (FAO, 2018). Seus produtos são largamente utilizados na produção de açúcar, álcool combustível e biodiesel.

A demanda mundial de açúcar é o principal incentivo para o cultivo de cana. A planta é responsável por 80% do açúcar produzido no mundo. O setor sucroalcooleiro brasileiro é de grande interesse de diversos países, principalmente pelo baixo custo de produção de açúcar e álcool. O etanol geralmente está disponível como um subproduto da produção de açúcar. Ele pode ser usado como uma alternativa de biocombustível à gasolina e é amplamente utilizado no Brasil. É uma alternativa para combustíveis fósseis e pode tornar-se o produto primário de processamento de cana-de-açúcar, ao invés do açúcar

(DAHLIA et al., 2009).

O cultivo da cana-de-açúcar é geralmente feito de forma extensiva, com o uso de monoculturas. As plantações ocupam vastas áreas contíguas e é necessária uma grande área plantada para justificar e manter produtiva a cadeia industrial à sua volta, as usinas de açúcar e de etanol. No entanto, os agricultores precisam conservar intocadas as áreas ao redor de mananciais de água, topo de montanhas e aclives acentuados, além de manter um percentual mínimo de mata nativa, que varia de acordo com a região, sendo 20% no Sudeste e até 90% na região amazônica (SILVA; FERREIRA, 2018).

No Brasil, a agroindústria da cana-de-açúcar tem adotado políticas de preservação ambiental que servem como métricas de exemplo internacional. Procura-se investir em inovações que acentuem a produção de seus derivados enquanto diminui o impacto sobre o meio ambiente. Devido à vasta área que as plantações de cana ocupam, se torna difícil, porém necessário, se prever fatores de crescimento ou declínio da produção das usinas, considerando aspectos como preservação do meio ambiente e metas de crescimento da produção.

Visando a maximizar a produção enquanto se procura minimizar os impactos ambientais, o agronegócio procura cada vez mais investir em tecnologias, pesquisas e inovações que auxiliem na análise de diversos dados coletados ao longo das safras. Tais dados devem servir como fonte para a obtenção de novas informações e para a tomada de decisões estratégicas.

Atualmente as usinas trabalham com a coleta de dados referentes a variáveis que possam vir a influenciar na qualidade de uma determinada safra. Essas variáveis são diversas, podendo-se destacar: temperatura, umidade, velocidade do vento, qualidade do solo, aparição de pragas, dentre outros. Com o objetivo de auxiliar no entendimento e obtenção de informação a partir destes dados pode-se utilizar técnicas de Mineração de Dados (MD).

O foco central da MD é o de como transformar dados armazenados em conhecimento, expresso em termos de formalismos de representação, tal como regras e relações entre dados (REZENDE, 2003). A MD é uma tecnologia que emergiu da interseção de três áreas: Estatística, Inteligência Artificial (IA) e Aprendizado de Máquina (AM). A MD possui o objetivo de extrair conhecimentos úteis (por exemplo, padrões) de dados complexos e vastos. As técnicas utilizadas para MD são de diferentes abordagens e suas aplicações dependem da natureza dos dados e do cenário do problema (KAMPFF, 2009).

Esses sistemas realizam uma análise de alto nível quanto a padrões ou tendências, mas também podem esmiuçar os dados para revelar mais detalhes, se necessário. Existem aplicações de MD para diferentes áreas funcionais das empresas, bem como para trabalhos científicos ou governamentais. Com o uso da MD, é possível descobrir informações relacionadas a associações, sequências, classificação, aglomeração e prognósticos (LAUDON KENNETH, 2011).

Muitas aplicações da MD são utilizadas para a descoberta de informações relativas a um meio específico ou a uma necessidade pontual, porém esta técnica também é utilizada para se prever comportamentos e acontecimentos de um determinado fenômeno. Outra abordagem da MD parte da aplicação em tempo real, provindos de fontes que fornecem informações em tempo constante. Essa abordagem vem ganhando mais força nos últimos tempos devidos a Internet das coisas, ou *Internet of Things* (IoT), e dos diversos sensores que vêm atrelados ao advento desta. A IoT é um conceito que visa a conexão de diversos dispositivos a internet, possibilitando uma captação de dados em tempo próximo ao real de diferentes fontes como relógios, tvs, sensores de temperatura, luz, humidade, som, dentre outros (SANTAELLA et al., 2013).

A tomada de decisão por descoberta da informação é muito útil para as usinas, mas o que elas mais investem e buscam atualmente são técnicas para que se possa prever uma determinada safra embasadas em modelagens e simulações de suas plantações. Este tipo de previsão envolve dados extraídos dos sensores da usina e análises de cenários reais, possibilitando que os gestores possam tomar decisões apoiados pelas informações de tempo e região de uma determinada plantação. É de interesse que diversos aspectos possam ser simulados e previstos, como pragas, secas, produção, alastramento do fogo nas queimadas, dentre outros.

Com isto, além da problemática já apresentada da necessidade de (I) extração de informações de um conjunto de dados complexos, surgem outras, como: (II) Previsão de acontecimentos; (III) Modelagem do ambiente e das variáveis as quais se deseja prever; (IV) Desenvolvimento de um modelo de simulação que se adapte a mudanças temporais e dê respostas para tomadas de decisão embasadas em dados em tempo próximo ao real.

Assim, percebe-se que é necessário fazer previsões e análises de comportamentos futuros a partir de objetos modelados. Contudo, a qualidade das respostas de um modelo qualquer proposto depende da precisão da estrutura computacional e dos dados que alimentam o modelo. A problemática levantada se trata de um sistema complexo, ou seja, um sistema que é composto por inúmeros elementos que interagem de modo que o comportamento agregado não pode ser inferido do comportamento das unidades constituintes isoladamente (FILHO; D'OTTAVIANO, 2000).

A literatura apresenta várias abstrações matemático-computacionais para auxiliar a modelagem e previsibilidade de comportamentos de objetos e sistemas complexos. Uma abordagem que vem sendo difundida por poder representar ambientes reais e ter uma adaptação com desempenho satisfatório para o consumo de dados em tempo real é o uso de Autômatos Celulares (AC) (MIRANDA et al., 2008).

AC representam sistemas dinâmicos, onde o tempo e o espaço são discretos (WOLFRAM, 2002). Eles vêm sendo utilizados na literatura como modelos para simulação de ambientes e objetos, incluindo fenômenos epidemiológicos, de forma matemático-computacional (FU, 2002)(ROUSSEAU et al., 1997). Autômatos são estruturas que con-

seguem representar cenários dinâmicos e são ideais para simular epidemias, queimadas e fenômenos aparentemente aleatórios.

Os AC geralmente seguem alguma regra definida, porém, uma abordagem muito utilizada para a modelagem de ambientes complexos é a que utiliza Autômatos Celulares Estocásticos (ACE). Os ACE são sistemas descritos por um conjunto de variáveis discretas em que os estados de cada célula são atualizados de forma síncrona e que obedecem a regras probabilísticas, dependendo apenas do estado de seus vizinhos no instante corrente (FERREIRA, 2009). Pelo fato de o estado corrente do autômato ser o único responsável pelo próximo estado global, a evolução de um ACE pode ser modelada por uma cadeia de Markov homogênea finita, e o resultado é a distribuição de probabilidades correspondente ao estado estacionário, obtida a longo prazo (LOZANO, 2017).

Outro método probabilístico muito utilizado para a modelagem de fenômenos é a aplicação de cadeias de Markov para simulações de cenários complexos. Pode-se modelar um problema em uma cadeia de Markov desde que se dependa apenas do estado em que o fenômeno se encontra e do estado a seguir para se determinar as predições. Cadeias de Markov são muito utilizadas em meteorologias, estudos econômicos, administrativos e sociais (BOLDRINI et al., 1980).

Este estudo utilizou como metodologia base o *Human-Centered Design* (HCD) do *Toolkit* HCD (IDEO, 2014). De acordo com IDEO (2014), o ato de projetar soluções inovadoras e relevantes, que atendam às necessidades, desejos e comportamentos dos contratantes, começa com o entendimento de suas necessidades e aspirações para o futuro. O HCD é fundamental para o projeto por possibilitar uma metodologia de imersão e compreensão da problemática e ideação para possíveis soluções viáveis. A ideia do *Toolkit* HCD consiste em transformar histórias coletadas em ideias implementáveis, com isso facilita-se a identificação de novas oportunidades e aumenta a velocidade e eficiência na criação de soluções inovadoras.

O projeto de pesquisa é focado em compreender e propor uma solução viável para uma problemática real e de impacto relevante para a agroindústria. Para isso foi realizado um estudo de caso com dados fornecidos pela Usina São José, pertencente ao grupo Cavalcanti Petribu. A São José Agroindustrial é uma das maiores produtoras de açúcar, etanol e energia elétrica de Pernambuco. Por ter um modelo industrial, tem um aproveitamento pleno da cana-de-açúcar, com uma produção comercializada nos mercados mundial e do Norte e Nordeste do Brasil. A usina é proprietária de uma área em torno de 28 mil hectares. Destes, 17 mil hectares são destinados à produção agrícola, e os demais correspondem às instalações da agroindústria e às áreas de proteção ambiental.

1.1 Motivação

A *Food and Agriculture Organization* (FAO) é uma agência das Nações Unidas que desenvolve e implementa metodologias e padrões para coleta, validação, processamento e análise de dados globais relativos a alimentos e a agricultura. A FAO destaca a importância da agricultura para o mundo, considerando-a como uma das bases para a civilização. Porém, faz um alerta que a monocultura industrial, o método de cultivo pelo qual a maior parte da oferta global de alimentos é cultivada, degrada a terra, reduz a resiliência ecológica e a diversidade, e requer uma enorme quantidade de combustíveis fósseis para se manter (BRUINSMA, 2017).

Sendo cientes das falhas existentes, produtores, pesquisadores e inovadores da monocultura devem buscar métodos e modelos alternativos para enriquecimento das terras, produção de alimentos saudáveis e métodos para a otimização das safras que ajudem a poupar o meio ambiente (DIAKOSAVVAS, 2012). Esse é um importante desafio de pesquisa a ser atacado e diferentes áreas do conhecimento estão trabalhando em cojunto para propor possíveis soluções.

A "*The future of work: regional perspectives*" (2018) é uma publicação realizada em parceria com bancos de diferentes continentes a qual avalia aspectos de inovação e do trabalho necessários para o impulsionamento futuro da economia mundial. A publicação de 2018 da *The future of work* retrata a importância da agroindústria para o cenário mundial e destaca que a confluência e o rápido desenvolvimento de uma ampla gama de novas tecnologias, como IA, robótica, impressão 3D e a IoT vem alavancando no mundo a quarta Revolução Industrial, possibilitando importantes ganhos de eficiência dentro das empresas e melhor qualidade de vida para os trabalhadores (BANK et al., 2018).

Seguindo este viés de pesquisa, pode-se encontrar muitas publicações recentes em meios científicos como a *Nature* (KNAPP; HEIJDEN, 2018) (HOLDEN et al., 2018) voltadas para a agricultura. Assim, percebe-se uma grande preocupação global com pesquisas que envolvam a otimização das produções agrícolas, bem como práticas sustentáveis de cultivos e qualidade de vida para os trabalhadores. Desta forma, a motivação inicial desta pesquisa parte da possibilidade de colaborar com possíveis soluções para um contexto que vem sendo discutido a nível global em que se tem o Brasil como uma das principais nações protagonistas do agronegócio.

Dois problemas de interesse da pesquisa foram observados na literatura e em campo durante a realização deste trabalho. (I) o controle de monoculturas e predição de safras é algo complexo de ser feito e estimado; (II) a adaptação e inserção das novas tecnologias e metodologias da indústria 4.0 ainda é algo nebuloso. Sendo assim, este trabalho visa a entrega de uma plataforma que possibilita a modelagem e simulação de monoculturas, já adaptado as tendências da 4ª revolução industrial. Porém, como foi visto, mapear plantações é algo tortuoso e caracterizado como um sistema complexo, ou seja, os componentes do sistema seguem interações não lineares entre si. Sistemas complexos são difíceis de

serem mapeados, mensurados e previstos.

O estudo das propriedades emergentes nos sistemas complexos possibilita a compreensão de seu comportamento global e da forma pela qual os componentes individuais se interagem para formar o todo (LOZANO, 2017). E por este motivo é tão importante para este trabalho a imersão na problemática e a escolha de modelos que possibilitem a representações destes sistemas complexos.

Outra grande motivação desta pesquisa parte de trabalhos e contribuições de (WOLFRAM, 2002) (MATA; COHN, 2007) (BALDUZZI; TONONI, 2008), dentre tantos outros pesquisadores da área que pregam a premissa de que, de fato, autômatos celulares são capazes de modelar qualquer fenômeno complexo. Por este motivo optou-se por utilizar como estrutura de modelagem os ACE.

Para a validação dos modelos propostos optou-se pela modelagem de monoculturas de cana-de-açúcar. Visto o fato de que o Brasil é o principal produtor da cana no mundo e pela relevância e leque de produtos derivados desta (FAO, 2018). Outro fator importante é o de que o Instituto Brasileiro de Geografia e Estatística (IBGE) afirma que

> Os produtos de cultura temporária de longa duração, como cana-de-açúcar e mandioca, cujo ciclo vegetativo ultrapassa 12 meses, e com período de colheita prolongado, devido a características de variedade, condições locais e finalidade a que se destina o produto colhido, necessitam de mecanismo complementar para o acompanhamento e estimativa da produção, sendo este um problema complexo. (IBGE, 2018).

O agronegócio não é importante apenas para o Brasil, mas para o mundo, sendo assim processos, inovações, pesquisas e investimentos têm que ser realizados e são necessários para garantir que este importante setor consiga crescer de forma sustentável, garantindo produtos saudáveis, preservação do meio ambiente, trabalho adequado e uma produção otimizada que atenda a demanda global.

1.2 Objetivos e Desafios da Pesquisa

O problema discutido neste trabalho é o de como modelar monoculturas visando simulações que ajudem na previsão e tomada de decisão para a otimização da produção e redução dos impactos ambientais por parte da agroindústria. Para propor uma possível solução para este problema, deve-se trilhar metodologias científicas que ajudem a compreender as reais necessidades dos *stakeholders*, provocando uma imersão na problemática enfrentada e que possibilite a coleta e identificação de dados e variáveis chaves. Com o conhecimento da problemática consolidado é preciso propor e testar possíveis soluções que atendam as demandas constatadas.

Diante do problema e desafios de pesquisa destacados, o objtivo deste estudo é imergir nas problemáticas enfrentadas nos processos produtivos da indústria sucroenergética, para

assim propor uma modelagem e simulação computacional que possibilite a realização de previsões e tomadas de decisão relativas as safras e qualidades de cultivo.

Para se chegar ao objetivo final desta pesquisa tem-se como objetivos específicos os seguintes tópicos:

- ❏ Imergir na problemática para a real compreensão das necessidades dos *stakeholders*, assim apresentando um panorama geral dos desafios identificados na literatura e em campo, investigando as principais variáveis de interesse a serem observadas no cultivo da cana-de-açúcar;
- ❏ Realizar uma revisão sistemática de métodos e algoritmos utilizados para promover a modelagem e simulação de plantações;
- ❏ Propor um modelo baseado em cadeias de Markov de Monte Carlo e outro modelo baseado em autômatos celulares estocásticos para a modelagem e simulação de cenários de cultivo e safras da cana-de-açúcar;
- ❏ Avaliar a solução proposta em um estudo de caso, cruzando e analisando resultados obtidos com dados históricos e de outras pesquisas. Com isto, espera-se analisar o comportamento dos modelos de autômatos celulares estocásticos e de cadeias de Markov de Monte Carlo;
- ❏ Prover a descoberta de conhecimento através da aplicação de diferentes técnicas estatísticas e de mineração nos dados coletados e analisados da agroindústria. Assim, identificando relações e informações importantes para os modelos e para a tomada de decisão;
- ❏ Contribuir com uma proposta de plataforma alinhada as novas necessidades da Indústria 4.0 e dos conceitos da agricultura de precisão, visando a otimização do cultivo e redução de impactos ambientais;
- ❏ Possibilitar que agrônomos e pesquisadores possam acompanhar as plantações remotamente e de forma georreferenciada, dado que essas possuam sensores enviando dados em tempo real, permitindo uma visualização das áreas onde determinados fatos podem acontecer;

1.3 Estrutura da Dissertação

Este trabalho está organizado em oito capítulos, dos quais:

- ❏ No Capítulo 1 encontra-se a introdução, mostrando os objetivos, motivações e contribuições;

- ❏ No Capítulo 2 são descritos os principais trabalhos relacionados a esta pesquisa, extraídos da revisão sistemática da literatura;
- ❏ No Capítulo 3 é descrita a fundamentação teórica do estudo, sendo dividida entre os principais tópicos de interesse teórico necessários para a compreensão deste trabalho;
- ❏ No Capítulo 4 são discutidas as metodologias e propostas utilizadas, apresentando melhor a problematização, arquiteturas, algoritmos e ferramentas de análises necessários para se trilhar esta pesquisa;
- ❏ O Capítulo 5 apresenta o modelo cadeia de Markov de Monte Carlo proposto, descrevendo suas características, exemplificando seu funcionamento e discutindo métricas, diagramas, algoritmos e observações constatadas;
- ❏ O Capítulo 6 apresenta o modelo de autômato celular estocástico proposto, e como no capítulo anterior, descreve-se suas características, exemplificando seu funcionamento e discutindo métricas, diagramas, algoritmos e observações constatadas;
- ❏ No Capítulo 7 é apresentado o estudo de caso, onde o projeto e modelos desenvolvidos foram testados e validados;
- ❏ No Capítulo 8 são descritas as conclusões obtidas com este estudo, as principais contribuições e os seus trabalhos futuros.
- ❏ Por fim, pode-se encontrar as referências utilizadas, os apêndices e anexos desta pesquisa.

Capítulo 2

Trabalhos Relacionados

Para embasamento teórico e científico desta pesquisa foi realizado um mapeamento sistemático de trabalhos e *softwares* relacionados aos temas abordados nesta dissertação. O objetivo é avaliar o atual estado da arte da agricultura de precisão, levantando as principais técnicas e ferramentas utilizadas na área.

Foram analisados trabalhos, pesquisas e aplicações reais de empresas referentes aos últimos dezoito anos. Diversos modelos e *softwares* foram identificados e estudados, mas pode-se perceber que ainda há muitas lacunas para a área, principalmente quando são abordados temas de otimização, melhorias genéticas, sustentabilidade e adaptação as novas tecnologias, como a IoT.

Um mapeamento sistemático é uma forma de identificar, avaliar e interpretar todas as pesquisas disponíveis relevantes para uma questão de pesquisa particular. Umas das razões para a realização de revisões sistemáticas é que ela resume as evidências existentes em relação a um tratamento, método ou tecnologia (KITCHENHAM, 2004).

Neste estudo foi realizado um mapeamento sistemático do ano 2000 à agosto de 2018 das seguintes bases de pesquisa: (I) Scielo; (II) ACM; (III) Springer; (IV) Google Scholar; (V) Scopus e (VI) IEEE Computer Society Digital Library. O objetivo desta análise de aproximadamente 18 anos de pesquisas é conhecer quais modelos, técnicas e ferramentas computacionais estão sendo aplicados na agricultura de precisão, buscando identificar lacunas na área e oportunidades de pesquisa e contribuição.

O mapeamento foi embasado no processo de condução de revisões e mapeamentos sistemáticos definido por Kitchenham (2004). O protocolo de mapeamento determina os métodos que serão usados para realizar um mapeamento sistemático específico, fazendo que este diminua a possibilidade de viés do pesquisador (KITCHENHAM, 2004).

O protocolo de pesquisa partiu da definição dos objetivos de pesquisa e para isso utilizou-se o paradigma GQM (*Goal-Question-Metric*) (BASILI; CALDIERA; ROMBACH, 1994), ilustrado na Tabela 1. Posterior a definição dos objetivos, foram definidas e descritas as questões de pesquisa (Tabela 2), refletidas através da representação do PICO (*problem or population; intervention; comparison; outcome*) (HUANG; LIN; DEMNER-

FUSHMAN, 2006). A partir destas definições escolheu-se as bases de pesquisa e as *strings* de busca.

Tabela 1 – GQM do Mapeamento Sistemático.

Analisar	
Com o propósito de	Identificar / Caracterizar / Avaliar
Em relação a	Simulações, Modelos, *Frameworks*, Mapeamentos e Validadores
Do ponto de vista dos	Pesquisas em modelagem e inteligência computacional
No contexto	Agricultura de precisão

Tabela 2 – Questões de Pesquisa.

1	Quais modelos, ferramentas, técnicas e *frameworks* vêm sendo utilizados para a modelagem e simulação de plantações ao longo dos anos?
2	Como a eficiência das pesquisas desenvolvidas tem sido avaliada?
3	Quais aspectos e variáveis são focos das ferramentas e propostas das pesquisas desenvolvidas?
4	Quais metodologias são utilizadas para se chegar a uma proposta de solução?
5	Quais os desafios e lacunas desta área de pesquisa?

A *String* de Busca utilizada foi:

(“Model” OR “Framework” OR “Simulation” OR “Tool” OR “Application” OR “Software” OR “Device” OR “IoT”) AND (“Agriculture” OR “Cultivation” OR “Planting” OR “Monoculture” OR “Sugarcane” OR “Harvest” OR “Vegetation” OR “Field”)

Para os resultados em português, foi utilizada a seguinte *string* de busca:

(“Modelo” OR “Framework” OR “Simulação” OR “Ferramenta” OR “Aplicativo” OR “Software” OR “Dispositivo” OR “IoT”) AND (“Agricultura” OR “Cultivo” OR “Plantio” OR “Monocultura”) OR “Cana-de-açúcar” OR “Colheita” OR “Vegetação” OR “Campo”)

O primeiro levantamento retornou 208 trabalhos. Kitchenham (2004) afirma que devem ser seguidos critérios de inclusão e exclusão para os artigos que são retornados pela *String* de busca. Sendo assim, foram definidos os critérios para inclusão, ou exclusão dos trabalhos obtidos na primeira etapa. Os critérios para a inclusão de artigos são:

- ❏ Estudos apresentando conceitos, modelos, teorias, discussões, relatos de experiência e revisões sistemáticas sobre agricultura de precisão.

Já os critérios para a exclusão de artigos são:

- ❏ Artigos não escritos em inglês ou português;
- ❏ Publicações não acessíveis na internet;
- ❏ Relatos de *Workshops* e *Keynotes*;
- ❏ Publicados antes dos anos 2000;
- ❏ Publicações que não satisfaçam a nenhum critério de inclusão.

Após definidos os critérios de inclusão e exclusão de artigos, defini-se o processo de seleção preliminar e final.

Processo de Seleção Preliminar (Filtro I):

Serão selecionados artigos que apresentem informações no título e/ou no *abstract* relacionado à questão de pesquisa principal. Para cada estudo incluído ou excluído será apresentado um critério (Inclusão ou Exclusão).

Processo de Seleção Final (Filtro II):

Como a leitura de duas informações (título e *abstract*) não é suficiente para identificar se o estudo é realmente relevante para a pesquisa realizada, torna-se necessário realizar a leitura completa dos estudos que restaram do Filtro I. Dessa forma, esta fase do mapeamento tem como objetivo fazer uma análise mais apurada dos estudos, identificando e extraindo dados também de acordo com os critérios de inclusão e exclusão descritos anteriormente. Para cada estudo incluído ou excluído será apresentado um critério (Inclusão ou Exclusão).

Após a realização das filtragens foram definidos critérios de qualidade com base na confiabilidade do método de avaliação aplicado no estudo. Utilizou-se o modelo de avaliação proposto por Bibi et al. (2014), considerando aspectos de precisão e relevância, como descritos a seguir:

Aspecto de avaliação: PRECISÃO

- ❏ Medidas: Strong (S), Medium (M), Weak (W)
- ❏ Pontuação: S = 1, M = 0,5 e W = 0

1. Como o contexto/ambiente de aplicação da pesquisa foi exposto?
2. Como a aplicação (coleta de dados) foi relatada?

Aspecto de avaliação: RELEVÂNCIA

- ❏ Medidas: Relevant (R), Not relevant (N)
- ❏ Pontuação: R = 1, N = 0

1. Qual a amostra analisada na pesquisa?
2. Qual o ambiente de aplicação da pesquisa?
3. Qual o método de pesquisa usado?

Por fim, foi possível identificar de que forma está acontecendo o avanço de pesquisas que envolvem computação na agronomia. Foi traçada uma linha do tempo com as principais contribuições e avanços de ferramentas, métodos e modelos computacionais

aplicados à agroindústria, revelando informações até então pouco investigadas e futuras oportunidades de pesquisa.

A partir do conjunto de 208 estudos analisados, 52 foram selecionados por serem mais relevantes ao tema pesquisado. A Tabela 3 ilustra a distribuição destas pesquisas ao longo dos anos. Pode-se perceber o aumento significativo na quantidade de trabalhos de interesse identificados nos últimos anos.

Tabela 3 – Distribuição de Trabalhos por Anos.

Ano	Publicações Selecionadas
2007	1
2010	4
2011	3
2012	2
2013	3
2014	3
2015	5
2016	8
2017	13
2018	10

O Google Trends é uma ferramenta do Google que mostra os termos mais populares buscados em um determinado período. A Figura 1 ilustra uma análise do termo "*smart/IoT agriculture*" na ferramenta e pode-se perceber uma comprovação da tendência observada nos trabalhos científicos destacados.

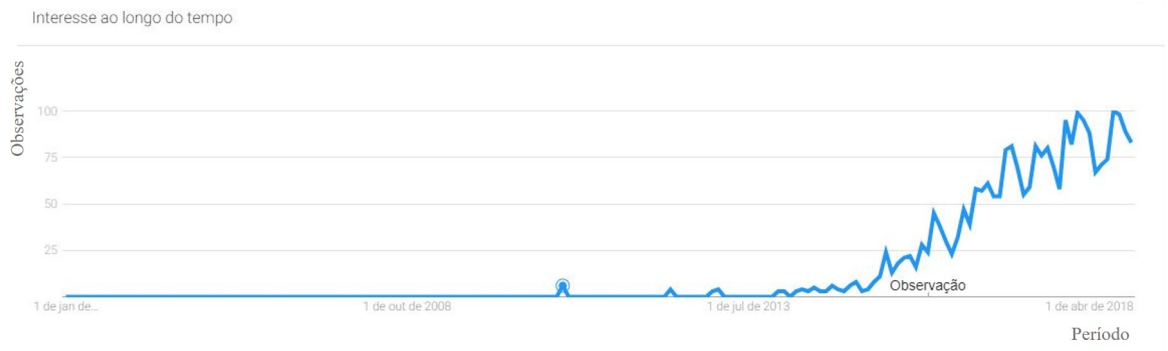


Figura 1 – Interesse ao longo do tempo por *smart/IoT agriculture*

Fonte: Adaptado de Google Trends (2018)

O mapeamento sistemático deu origem a um artigo que descreve o seu detalhamento e os seus principais resultados. Entre os principais resultados obtidos, percebe-se o crescimento de pesquisas na área, e a relevância e impacto que a agricultura de precisão vem tendo no cenário mundial. A computação vem sendo cada vez mais utilizada para a otimização de métodos de cultivos, mas ainda há brechas para se lidar com novos conceitos e métodos tecnológicos, como a IoT, *blockchain*, dentre outros que cercam a ideia da indústria 4.0. Busca-se uma agricultura sustentável, saudável, otimizada e segura para a população.

O mapeamento levantou trabalhos referentes a diferentes temáticas/abordagens tecnológicas, visando ao foco desta pesquisa. Nesta seção serão apresentados os principais trabalhos que abordam a utilização de AC, técnicas de MD, uso de cadeias de Markov e simulações em ACE, dentro da agroindústria.

2.1 *Smart Agriculture* e Modelagem Computacional de Monoculturas

Há muitos estudos e técnicas para a promoção da *Smart Agriculture*, principalmente por ser uma grande área que envolve conceitos de química, biologia, engenharia, dentre outras. Focando em pesquisas que envolvam a aplicação de computação na *Smart Agriculture* também é possível identificar muitos trabalhos interessantes, envolvendo recursos de IoT, IA e modelagem computacional.

Ao se observar pesquisas e trabalhos mais antigos, percebe-se que a computação era normalmente aplicada como facilitadora para o processamento de modelos matemáticos. O trabalho de Martin e Ek (1984) faz uma comparação entre vários modelos matemáticos para o crescimento de pinheiros, através de regressões não lineares. Os autores concluem que o modelo empírico é o mais preciso para projeções de crescimento de diâmetro em plantios manejados. No entanto, discorrem que o modelo semi-empírico possivelmente será mais preciso quando não se tem tanto conhecimento sobre a plantação analisada. As conclusões do trabalho de Martin e Ek (1984) são relevantes para o que se existe ainda hoje em diferentes plantações, onde normalmente alguns especialistas da área são os responsáveis por empiricamente realizarem previsões para tomadas de decisão. Porém, o advento do poder de processamento computacional e novas formas para coleta de dados trazem novos paradigmas para modelagens e simulações em plantações.

Suma et al. (2017) desenvolveram um estudo na Índia para a automação da coleta de dados agrícolas utilizando a IoT. Foi realizado um mapeamento de sensores e atuadores para o monitoramento remoto de plantações baseado em GPS, detecção de umidade, temperatura, intrusão de animais, segurança, umidade foliar e irrigação. O projeto de Suma et al. (2017) também utiliza redes de sensores sem fio para observar continuamente as propriedades do solo e os fatores ambientais, possibilitando o controle destes parâmetros através de qualquer dispositivo com acesso a internet.

Rao et al. (2018) também fazem uso da IoT, só que focada na automação da irrigação visando a otimização do uso de recursos hídricos. Estes dois trabalhos concluem que a utilização da IoT agrega valores significativos para a automação de processos e otimização de recursos das plantações.

Lin et al. (2018) propõem um sistema confiável, auto-organizado, aberto e ecológico de rastreabilidade de alimentos baseado em *blockchain* e IoT. A plataforma proposta envolve todas as partes de um ecossistema agrícola inteligente, deste o plantio até a venda. O

trabalho utiliza dispositivos IoT para substituir a coleta de dados manual e o *blockchain* para certificar a confiabilidade de todas as transações realizadas durante o ciclo de plantio a venda de alimentos. O grande foco da proposta feita por Lin et al. (2018) é reduzir a intervenção humana e idealizar uma forma para usar a tecnologia de contratos inteligentes para ajudar os envolvidos a encontrarem problemas e processá-los em tempo hábil.

Dois trabalhos muito interessantes e norteadores para a identificação de lacunas na área da agricultura de precisão foram os de Mehta e Patel (2016) e Mekala e Viswanathan (2017), nos quais foi realizado um levantamento do estado da arte buscando identificar oportunidades e desafios. Eles apresentam aplicações típicas de IoT na agricultura usando a computação em nuvem como *backbone*.

O trabalho de Mehta e Patel (2016) destaca o potencial de crescimento da indústria de embarcados e IoT, apontando para a necessidade de tornar essas tecnologias comuns e úteis para o cotidiano das pessoas. Espera-se que a IoT conecte 28 bilhões de diferentes dispositivos até 2020, com a agricultura sendo um dos setores privilegiados por esses avanços. Questões relativas à agricultura são fundamentais para o mundo, pois ela abastece as mesas das famílias e modernizar os métodos tradicionais de agricultura é fundamental diante do crescimento populacional e da alta demanda por este setor.

O levantamento de Mekala e Viswanathan (2017) destaca que a IoT é uma tecnologia revolucionária que representa o futuro da computação e das comunicações. A maioria das pessoas no mundo depende da agricultura. Por essa razão, são necessárias tecnologias inteligentes e eficientes para sofisticar os métodos tradicionais de agricultura. O uso de tecnologias modernas pode controlar o desempenho de custo, manutenção e monitoramento das plantações. Imagens de satélite e aéreas desempenham um papel vital na agricultura moderna. A rede de monitoramento de sensores de agricultura de precisão é usada para medir informações relacionadas ao cultivo, como temperatura, umidade, PH do solo, níveis de nutrição do solo, nível de água, dentre outros. Esses dados são disponibilizados remotamente para os agricultores.

Portanto, estes dois trabalhos convergem para a ideia de que os principais desafios para a área são relativos a: (I) Integração entre diferentes *hardwares* e *softwares*; (II) Conseguir unir conceitos da computação e da agronomia; (III) Segurança da informação; (IV) Modelar computacionalmente cenários reais; (V) Tempo de processamento para realizar simulações precisas; (VI) Identificação de variáveis e fenômenos de interesse e correlatos.

O trabalho de Kapoor et al. (2016) mescla conceitos de IoT com os de processamento de imagem, destacando que estes já foram aplicados individualmente no contexto da agricultura e que alcançaram certo grau de sucesso, entretanto, a combinação destas tecnologias era inexistente. Os autores descrevem uma abordagem para combinar IoT e processamento de imagem, a fim de determinar o fator ambiental ou fator artificial (pesticidas / fertilizantes) que está impedindo o crescimento das plantas. Utilizando uma rede de sensores que fazem as leituras dos fatores ambientais cruciais e das imagens de

folhas, é feito um processamento dos dados utilizando o *software* MATLAB e com a ajuda da análise do histograma é possível se chegar a resultados conclusivos.

Roy et al. (2017) propõem uma plataforma para disponibilização dos dados coletados a partir da IoT em nuvem, apresentando o AgroTick, um sistema híbrido para agricultura inteligente. O AgroTick é um sistema baseado em IoT suportado com interface móvel e projetado usando módulos de tecnologia como computação em nuvem, *firmware* incorporado, unidade de *hardware* e análise de *big data*. O AgroTick é arquitetado e projetado para melhorar a eficiência da agricultura, construir uma rede agrícola bem conectada e criar uma plataforma de compartilhamento de conhecimento para os agricultores.

A pesquisa de Jesus, Silva e Rocha (2015) destaca que a economia brasileira tem sido impulsionada pelo setor agrícola e que nesse contexto é preciso realizar ações preventivas que venham efetuar o controle e monitoramento de pragas e doenças para garantir o sucesso e qualidade das safras. Esta pesquisa aborda o desenvolvimento e uso de uma aeronave não tripulada (drone) autônoma para captura de fotos de plantações de soja, bem como um *software* de processamento destas imagens que possibilita a detecção de possíveis pragas e doenças na plantação.

Alves, Souza e Marques (2005) fazem uma comparação entre metodologias para determinar o potencial de erosão do solo. A Equação Universal de Perda do Solo (EUPS) é comparada com uma implementação da Lógica Fuzzy de duas variáveis. Os dados de entrada são mapas de declividade e cobertura da terra. Ambos os modelos (Fuzzy e EUPS de duas variáveis) foram aplicados em uma área de testes. Os resultados obtidos com os dois modelos foram comparáveis em relação à distribuição espacial e à porcentagem da área da bacia ocupada. O modelo Fuzzy de duas variáveis conseguiu estimar áreas com potencial de erosão de moderado a alto, usando um número reduzido de variáveis. Tal característica torna o modelo Fuzzy de duas variáveis adequado particularmente para regiões caracterizadas por recursos limitados, informação insuficiente e uso insustentável do solo.

Silva e Rakocevic (2011) apresentam a construção de maquetes de plantas de erva-mate em vários estágios de desenvolvimento com o uso do programa InterpolMate e a possibilidade de computar a fotossíntese a partir de estrutura interpolada. Os valores estimados a partir das maquetes, incluindo dimorfismo sexual e fotossíntese foliar, são muito próximos aos observados em campo. Contudo, essa semelhança foi limitada às reconstruções que incluíram unidades de crescimento de conjuntos de dados originais. A modelagem da dinâmica de crescimento possibilita estimativas de fotossíntese de plantas inteiras da erva-mate, o que é dificilmente mensurável no campo. O estudo conclui que o *software* InterpolMate é eficiente na construção de maquetes de erva-mate.

A redução sistemática de florestas naturais, provenientes de sucessivos incêndios, tem estimulado o desenvolvimento de mecanismos de prevenção, controle e combate ao fogo. Juvanhol (2014) apresenta uma pesquisa focada em determinar as áreas de Risco de

Incêndios Florestais (RIF) utilizando informações do uso e ocupação da terra, declividade do terreno, orientação do relevo e proximidade a estradas e residências. Com o auxílio de técnicas de Sistemas de Informações Geográficas (SIG) determinou-se a influência de cada variável ao RIF por meio da modelagem Fuzzy no aplicativo computacional ArcGIS/ArcINFO 10.0. O modelo desenvolvido por Juvanhol (2014) se mostrou adequado na avaliação dos impactos de diferentes variáveis sobre o risco de incêndio.

O objetivo do estudo de Junior et al. (2015) é avaliar a variação da cobertura vegetal (áreas de floresta e manguezal) utilizando geotecnologias. Com a finalidade de se obter os mapas temáticos, foram realizadas visitas a campo para coleta de diversas informações, identificação e mapeamento acerca das principais atividades desenvolvidas na área, com auxílio aparelhos de GPS, câmeras fotográficas digitais e imagens de satélites da série Landsat (5TM e ETM 7), fornecidas pelo Instituto Nacional de Pesquisas Espaciais (INPE). A partir das análises, é possível verificar diminuições na cobertura vegetal das florestas e de manguezais, discutir sobre fatores que estão levando a esta diminuição e a realização de estimações futuras para essas áreas. Os autores, por fim, ressaltam que pesquisas que envolvem o georreferenciamento são muito importantes para a preservação e controle ambiental de precisão e que mais projetos são necessários para refinar a área.

Modelos Fuzzy, processamento de imagem e muitas outras técnicas podem ser utilizadas para gerar modelos ou promover predições e informações para tomadas de decisão na agricultura. Como é mostrado nas próximas seções, muitos trabalhos estão utilizando AC e cadeias de Markov para modelagens de ambientes complexos. A pesquisa desenvolvida nesta dissertação visa ao desenvolvimento de um modelo embasado em ACE, que pode ser alimentado por dados oriundos de pesquisadores ou de dispositivos IoT, para o acompanhamento georreferenciado e realização de simulações de monoculturas.

2.2 Cadeias de Markov Aplicadas para a Modelagem de Ambientes e Fenômenos

Cadeias de Markov são importantes representações de estados e transições estocásticos para a modelagem e simulação de diferentes fenômenos de tempo discreto e contínuo. Diversos trabalhos utilizam cadeias de Markov como estruturas para a representação de fenômenos que precisam de algum tipo de predição.

Os economistas agrícolas estão frequentemente interessados em caracterizar como os processos econômicos mudaram ao longo do tempo, bem como que caminhos eles provavelmente tomarão em períodos futuros. Dado esse interesse, vários modelos vêm sendo estudados e aplicados ao longo dos anos. A exemplo desta evolução, o trabalho de Judge e Swanson (1962) visam a métodos de análise que permitam atingir esses objetivos agrícolas e que sejam simples de aplicar já a mais de 50 anos atrás. É discutido e apresentado o conceito de um processo em cadeia de Markov para lidar com problemas onde dados

detalhados de longos períodos são analisados para se obter estimativas. O trabalho lida com amostras de dados de empresas produtoras de suínos no centro de Illinois, obtendo resultados satisfatórios.

Os sistemas determinísticos têm seu comportamento conhecido ou determinável ao longo do tempo, o que não se pode dizer dos sistemas probabilísticos. Estes são baseados em probabilidades de acontecimentos e requerem uma ferramenta auxiliar para que sua análise possa ser realizada, como uma cadeia de Markov. O trabalho de Marcos (2015) propõe um modelo de cadeia de Markov através de um problema de manejo de pragas de lavouras cafeeiras. Foram abordadas duas dimensões: incidência e pulverização. Os resultados de três anos consecutivos foram levantados, abordando percentuais de transição recorrente ao estado sem incidência e sem pulverização, bem como o percentual de estados sem incidência e sem pulverização. Também foram levantados os setores com incidências e pulverizações máximas e mínimas nos períodos. A cadeia de Markov demonstrou ser uma ferramenta eficiente no monitoramento de pragas e análise global do sistema, o que é essencial para o desenvolvimento da agricultura de precisão.

A pesquisa de Silveira Júnior (2014) visa a aplicação de cadeias de Markov para auxílio no controle biológico da planta aquática *Eichhornia azurea* por meio da inserção do inseto predador *Thrypticus sp.* Dados obtidos de uma pesquisa desenvolvida pelo autor entre 2002 e 2003 foram incorporados no processo de modelagem. Simulações de diferentes cenários foram realizadas, supondo infestação pela planta em represas interligadas e os resultados mostraram a quantidade de insetos necessários para controle em cada ciclo. O uso da modelagem matemática e das cadeias de Markov permite a implementação de vários testes.

López et al. (2001) apresentam um trabalho relativo à simulação da cobertura e uso de terras em áreas mapeadas (nos últimos 35 anos) nos arredores de uma cidade em crescimento no México. São utilizadas fotografias, técnicas de áreas retificadas e SIG. A mudança na cobertura de terra foi estimada para os 20 anos seguintes usando cadeias de Markov e análises de regressão. O estudo explorou as relações entre crescimento urbano e mudança de paisagem, e entre crescimento urbano e crescimento populacional. A simulação apresentou resultados positivos, porém, pode-se perceber que gramados e arbustos são categorias instáveis para a predição. O uso mais poderoso das matrizes de transição de Markov parece estar no nível descritivo, e não no preditivo. A regressão linear entre o crescimento urbano e populacional ofereceu uma previsão mais robusta do crescimento urbano.

Para combinar a proteção e a utilização dos recursos florestais nos trópicos, a compreensão da dinâmica florestal é essencial. Em silvicultura, a dinâmica florestal pode ser traduzida como a compreensão das taxas de recrutamento, mortalidade e incremento de biomassa ao longo do tempo. Teixeira et al.(2007) estimam essas taxas com base em medições realizadas em 2000 e 2004. A pesquisa trata da dinâmica florestal de uma floresta

intocada baseada na matriz de transição probabilística (cadeia de Markov de primeira ordem). O objetivo principal é relatar mudanças de 4 anos (2000 a 2004) na estrutura da floresta. Os resultados alcançados indicaram que a abordagem da cadeia de Markov é uma ferramenta confiável para projetar a dinâmica da floresta em curto prazo.

O trabalho de Keller Filho, Junior e Lima (2006) tem como objetivo verificar se as ocorrências de dias secos e chuvosos são condicionalmente dependentes da sequência dos três dias secos e chuvosos anteriores, numa zona pluviometricamente homogênea, por meio da cadeia não-homogênea de Markov de terceira ordem. Os resultados mostraram que as probabilidades diárias de transição podem ser adequadamente estimadas, com base em dados agregados bimestralmente, seguidas de interpolação por meio de funções sinusoidais. Além disso, evidenciou-se que, naquela zona, as ocorrências diárias de chuva são condicionalmente dependentes da sequência de dias secos e chuvosos nos três dias anteriores. A cadeia não-homogênea de Markov de terceira ordem é um importante instrumento para a análise da dependência entre as sequências de dias secos e chuvosos em determinadas regiões

O primeiro modelo desenvolvido neste trabalho para a realização das simulações das monoculturas é embasado em cadeias de Markov, em que fases do cultivo são divididas em estados da cadeia e as probabilidades de transição e ciclos são derivados da MD e de ajustes por especialistas. Os resultados alcançados com o modelo cadeia de Markov são utilizados como base para o desenvolvimento e refino do segundo modelo, que utiliza a ideia de ACE.

2.3 Autômatos Celulares Aplicados ao Contexto de Modelagem de Ambientes e Fenômenos

Modelos computacionais baseados no paradigma de AC foram concebidos recentemente para a simulação de mudanças de uso e cobertura da terra, a exemplo de expansão agrícola, processos de desmatamento, crescimento urbano, entre outros. O trabalho de Macedo, Almeida e Santos (2018) tem como objetivo aplicar e avaliar um modelo de AC para simular mudanças de uso e cobertura da terra vinculadas à expansão agrícola. As simulações foram realizadas entre o período de 2003 a 2012, sendo as simulações de 2002 a 2008 retrospectivas, e as simulações de 2009 a 2012 prospectivas. A aplicação do modelo revelou que houve expansão agrícola durante o período analisado, em detrimento de ambientes nativos, tais como matas e campos serranos. Os resultados alcançados revelam o potencial da aplicação dessa classe de modelos para o estudo e a predição de mudanças de cobertura e uso da terra.

Peixoto, Barros e Bassanezi (2003) realizaram uma pesquisa relativa à Morte Súbita dos Citros (MSC). A MSC é uma doença de causa ainda desconhecida, que tem afetado e matado plantas em alguns lugares do mundo, inclusive no Brasil, no sul do Triângulo

Mineiro e norte do Estado de São Paulo. A doença recebeu este nome devido a rapidez com que as árvores cítricas morrem após a adquirirem. Estes autores adotaram um modelo de AC para representar o espalhamento da doença, que incorpore aspectos de incertezas, utilizando controladores Fuzzy para modelar as variáveis incertas. Através do modelo criado foi feito o estudo da evolução temporal e espacial da doença. Concluiu-se que o padrão de espalhamento da doença é não local e depende da força do vento, e o progresso temporal da doença em um talhão de citros, sendo mais acentuado na primavera, obedece a um crescimento logístico a cada ano.

Autômatos Celulares Probabilísticos (ACP) são utilizados como uma abordagem para modelagem e simulação de incêndios de vegetação. O trabalho de Almeida et al. (2015) apresenta um formalismo de modelagem que possibilita estabelecer uma relação explícita dos parâmetros do modelo com dados meteorológicos e espaciais obtidos por SIG. Um amplo acervo de dados espaciais que inclui o inventário histórico de incêndios mapeados, mapa de vegetação, modelo digital de elevação, malha de drenagem, malha de aceiros, somados a dados meteorológicos obtidos por estações de coleta, deram subsídios para a compreensão e a modelagem do comportamento do fogo em uma unidade de conservação específica. A metodologia de ajuste desenvolvida mostra a boa capacidade de o modelo ser ajustado para simular incêndios reais, com certo nível de precisão em termos de extensão do incêndio e tempo de duração.

Lozano (2017) faz um estudo relativo a sistemas biológicos complexos, ou seja, o entendimento só é possível observando interações entre os componentes individuais e das propriedades emergentes. Esses sistemas são altamente adaptativos, dinâmicos e evoluem ao longo do tempo por meio do processamento da informação, se assemelhando ao cenário objetivo de plantações e cultivo. O objetivo deste trabalho é identificar modelos que possam ser aplicados a sistemas biológicos. Foram considerados AC binários unidimensionais e regras elementares. Na evolução do autômato foi introduzida incerteza no nível celular, de forma que cada célula tem a possibilidade de desobedecer a regra em uso e alterar seu estado de forma distinta daquela determinada pela regra. A análise dos resultados sugere que, ao tentar reduzir a incerteza que emerge das interações, os componentes do sistema geram informação. Esses resultados podem ser utilizados, metaforicamente, na representação de processos biológicos complexos, tais como a resposta imune (sistema imunitário) e o aparecimento de estados conscientes (sistema neuronal).

O estado de São Paulo é o principal produtor de cana-de-açúcar no Brasil, e a agroindústria sucroalcooleira está em expansão, substituindo outras culturas e penetrando em regiões que não eram tradicionalmente ligadas à produção canavieira. Visando a este cenário a abordagem proposta por Adami (2011) busca integrar modelos de AC, SIG e métodos de avaliação multicritério para simular a dinâmica espacial da cana-de-açúcar no território do escritório de desenvolvimento regional de Araçatuba (SP) entre o período de 2001 a 2008. A partir da opinião dos gestores e especialistas envolvidos, um modelo

AC foi empregado para simular o crescimento da cultura canavieira na área de estudos. Os resultados foram comparados àqueles obtidos pela aplicação de métodos tradicionais, como cadeias de Markov, e classificações técnicas, exemplificada pela capacidade de usos das terras. Todos estes processos embasaram a aplicação de modelos AC.

Os resultados deste estudo de Adami (2011) demonstraram que os métodos propostos, ao incorporar as preferências dos atores e dos especialistas, apresentaram desempenho superior na simulação da dinâmica espacial da cultura canavieira. Foram obtidos valores mais elevados nos índices Kappa, acurácia total e Lee-Salleem, menores erros de omissão e comissão para a cultura canavieira, além de aproximações maiores em parâmetros simples de análise da paisagem quando comparados aos métodos tradicionais. Os procedimentos metodológicos propostos permitem simular a expansão da cana-de-açúcar e dimensionar seus efeitos sobre os usos das terras em nível regional.

Gradativamente, os agricultores têm percebido que a tomada de decisões e uso de sistemas inteligentes é mais que uma simples tendência, mas uma ação de sobrevivência e obrigação, justificada pela exigência cada vez maior do mercado consumidor quanto às questões de segurança alimentar, respeito ao meio ambiente e saúde do trabalhador rural. No entanto, existem poucos sistemas que atendem aos pequenos produtores. Dentro deste cenário, Oliveira e Vianna (2015) desenvolvem um trabalho focado para modelagem e simulações de pequenas plantações de tomate. A abordagem aplicada é embasada em uma modelagem computacional em AC e busca a melhoria da qualidade das culturas de tomate. A pesquisa apresenta o Sistomate, um sistema de apoio à tomada de decisões relacionadas à produção de tomates.

O estudo de Marko, Zulkarnain e Kusratmoko (2016) tem como objetivo prever mudanças na cobertura da terra usando o acoplamento de cadeias de Markov e AC. Cadeias de Markov tem uma boa capacidade de prever a probabilidade de mudança estatisticamente, enquanto AC são um método poderoso na leitura dos padrões espaciais de mudança. Os dados da cobertura temporal do solo foram obtidos por imagens de satélite de sensoriamento remoto. Além disso, este estudo também utilizou análise multicritério para determinar qual fator determinante poderia estimular as mudanças, como proximidade, elevação e inclinação. O acoplamento desses dois métodos poderia fornecer um melhor modelo de previsão, em vez de apenas usá-lo separadamente. O modelo de predição foi validado usando os dados de cobertura da terra existentes em 2015 e mostrou um coeficiente Kappa satisfatório.

O trabalho de Gidey et al. (2017) tem como objetivo predizer e analisar os cenários futuros de cobertura do solo utilizando AC e cadeia de Markov (Modelo AC Markov), levando em consideração fatores físicos e socioeconômicos. Dados históricos de mudança da cobertura do solo foram usados como base. Tanto as regras de transição quanto a matriz da área de transição foram produzidas quantitativamente usando o modelo AC Markov. Os fatores físicos e socioeconômicos foram padronizados usando Fuzzy e, em

seguida, utilizou-se *Multi-Criteria Evaluation* (MCE) para produzir a imagem de adequação de transição. O modelo AC Markov foi aplicado como um filtro de contiguidade padrão de 5 por 5 para prever a condição da cobertura do solo usando o *software TerrSet Geospatial Modeling and Monitoring System*. O resultado da correlação de Pearson entre o tipo da cobertura do solo histórico e o predito indicou que existem relações positivas, fortemente correlacionadas e estatisticamente significativas (r = 0,981, p = 0,000).

Devido a presença de incertezas e aleatoriedades no mundo real, é difícil simular a mudança do uso das terras com exatidão. Para resolver o problema da incerteza e aleatoriedade temporal e espacial, Huang (2015) propõe um modelo para simular essas mudanças do uso da terra com base na cadeia de Markov e em AC. As regras de transição do modelo foram definidas primeiro por condições globalmente restritas, condições localmente restritas e uma variável aleatória. E então os padrões de uso da terra e as mudanças foram obtidos a partir de imagens classificadas do Landsat TM. Uma matriz de transição representando o espaço-temporal foi construída a partir das imagens classificadas e foi aplicada ao modelo proposto para simular as mudanças. Os resultados do experimento mostram a validade e viabilidade do modelo baseado em Markov-AC para simular mudanças de terrenos urbanos.

O IDRISI (2016) é um *software* de SIG e processamento de imagens com ênfase em funções de análise. Ele reúne um conjunto de módulos que abrangem um grande número de operações analíticas, desde ferramentas básicas para cálculo de distância, até ferramentas mais sofisticadas para análises complexas, por exemplo, o *Land Change Modeler* e o *Earth Trends Modeler*. O Markov-AC é um dos principais modelos estudados pela equipe do IDRISI e vem sendo aplicado em pesquisas que utilizam a ferramenta em diferentes países, alcançando resultados satisfatórios.

O grande diferencial do modelo final desenvolvido nesta dissertação é a entrega de uma plataforma que foi estudada e estruturada para se integrar a cenários reais de sensores e atuadores da agricultura de precisão com o arcabouço de modelagem e predição em nuvem dos AC, sendo possível mapear e monitorar a plantação remotamente. O modelo visa a um baixo custo computacional para o processamento de dados em tempo próximo ao real para a tomada de decisões e elaboração de planejamentos para diferentes cultivos. Os demais trabalhos levantados utilizam algum processamento de imagem de satélite ou drone para realizar as simulações e essas geralmente são realizadas com dados históricos não recentes.

Capítulo 3

Fundamentação Teórica

Esta pesquisa abrangeu diferentes conhecimentos da ciência da computação e de outras grandes áreas de pesquisa, passando por MD, inteligência e modelagem computacional, IoT, agricultura de precisão, indústria 4.0, dentre outros, nos quais foram levantados diversos trabalhos científicos para fundamentar o desenvolvimento deste projeto. Esta seção será dedicada para os principais conceitos e teorias relacionados a pesquisa realizada.

3.1 Cultivo e Utilização da Cana-de-açúcar

A cana-de-açúcar originou-se na Nova-Guiné e foi transportada para o sul da Ásia, onde foi usada, primeiramente em forma de xarope. O primeiro relato do açúcar em sua forma sólida, foi na Pérsia. Pertencente à família *Gramineae* (*Poaceae*), a cana-de-açúcar de gênero *Saccharum*, composto principalmente pela espécie *Saccharum officinarum*, é conhecida por cana-nobre pois apresenta elevado teor de açúcar (SEGATO et al., 2006). O colmo da cana concede a matéria-prima para a fabricação de produtos como açúcar, rapadura, cachaça e álcool. A ponta, ou olho da cana, pode ser aproveitado na alimentação de animais e os resíduos da cana decorrente da fabricação do açúcar e do álcool podem ser usados na recuperação dos solos (CENTEC, 2004).

Dado o leque de produtos possíveis originados a partir da cana-de-açúcar, diferentes países buscam investir em plantações e indústrias sucroenergéticas para poderem fazer uso dos recursos obtidos da planta. Apesar do interesse em importações e exportações do produto, são precisas condições ambientais específicas para o seu cultivo.

Aproximadamente metade da cana-de-açúcar que se produz no mundo é de responsabilidade de algumas nações das Américas: Brasil, Cuba, México e EUA (TORQUATO, 2006). Pelas condições de clima e solo encontradas no Brasil, apresentando excelentes condições para o cultivo e implantação da monocultura da cana-de-açúcar, o país se tornou líder mundial na produção deste *commoditie*. A Figura 2 ilustra as principais nações produtoras de cana no mundo (FAO, 2018).

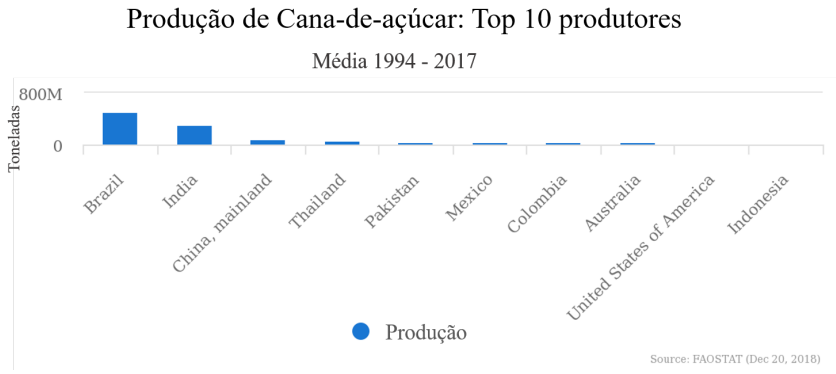


Figura 2 – Maiores Produtores de cana-de-açúcar

Fonte: Adaptado de FAO (2018)

O processo produtivo de um canavial visa três fatores básicos, segundo Câmara (1993):

- ❏ Produtividade: alto rendimento agrícola de colmos industrializáveis;
- ❏ Qualidade: riqueza em açúcar dos colmos industrializáveis;
- ❏ Longevidade do canavial: visa o aumento do número de cortes, considerando um prazo maior de tempo entre as reformas do canavial.

3.1.1 Botânica e fisiologia da cana-de-açúcar

A cana-de-açúcar desenvolve-se em forma de touceira. A parte aérea é composta por colmos (caule), folhas, inflorescência (conjunto de flores) e frutos, enquanto a subterrânea é composta por raízes e rizomas. Os rizomas são constituídos por nós, espaço entrenós e gemas, as quais são responsáveis pela formação dos perfilhos, como pode ser observado na Figura 3. As novas touceiras chamadas soca e ressoca se originam dos rizomas que rebrotam depois da colheita (SEGATO et al., 2006).

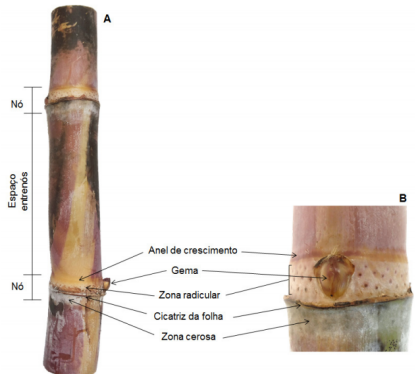


Figura 3 – A)Tolete e B) Nó do tolete de cana-de-açúcar e suas partes.

Fonte: Thomas (2016)

Como a maioria das gramíneas, a cana-de-açúcar é uma planta com ciclo C4, assim chamada por apresentar maior taxa fotossintética e de eficiência na utilização e absorção de CO2 (gás carbônico) da atmosfera. As plantas que pertencem a esse ciclo fotossintético conseguem aproveitar e resistir melhor às condições ambientais e produzir mais açúcares mesmo diante de uma deficiência hídrica prolongada. Essas plantas suportam temperaturas até 35ºC e alta incidência luminosa, sem interferir drasticamente nas atividades fotossintéticas (KIMATI et al., 1997).

A cana-de-açúcar é habituada às condições de alta luminosidade, altas temperaturas e relativo estresse hídrico, mesmo a cultura necessitando de grandes quantidades de água para suprir as suas necessidades hídricas, já que somente 30% do seu peso é representado pela matéria seca e, 70% pela água (SEGATO et al., 2006). Ou seja, em algumas situações a cana precisará de maior quantidade de água para seu desenvolvimento, e em alguns casos, pouca água poderá ser eficiente para a qualidade da comercialização dos colmos.

Portanto, a cana é uma excelente espécie para cultivo em regiões tropicais. No entanto, é necessário o conhecimento do ciclo da cultura para melhor manejá-la, pois sabe-se que qualquer produção agrícola que objetiva à produtividade econômica está associada na interação de três fatores: a planta, o ambiente de produção e o manejo (RODRIGUES, 2017).

3.1.2 Plantio e fases de crescimento da cana

Quanto as fases de crescimento da cana-de-açúcar, pode-se classificar o desenvolvimento vegetativo e reprodutivo da cultura em dez estágios fenológicos. Porém, em termos de produção agrícola (produção de etanol e açúcar), essa classificação fenológica não é muito utilizada, pois os estados de desenvolvimento reprodutivo não são desejáveis (florescimento). Dessa maneira, classifica-se o crescimento em quatro fases: brotação, perfilhamento, crescimento vegetativo e maturação, como ilustrado na Figura 4.

Figura 4 – Fases de desenvolvimentos da cana-de-açúcar.

Fonte: Adaptado de YARA (2018)

O ciclo da cana plantada pela primeira vez, originada de muda ou tolete que receberá o

primeiro corte, chama-se ciclo da cana-planta. Após o corte da cana-planta, há um novo ciclo de aproximadamente 12 meses, chamado ciclo das soqueiras ou cana-soca, como ilustrado na Figura 5 (GLASSOP; RAE; BONNETT, 2014).

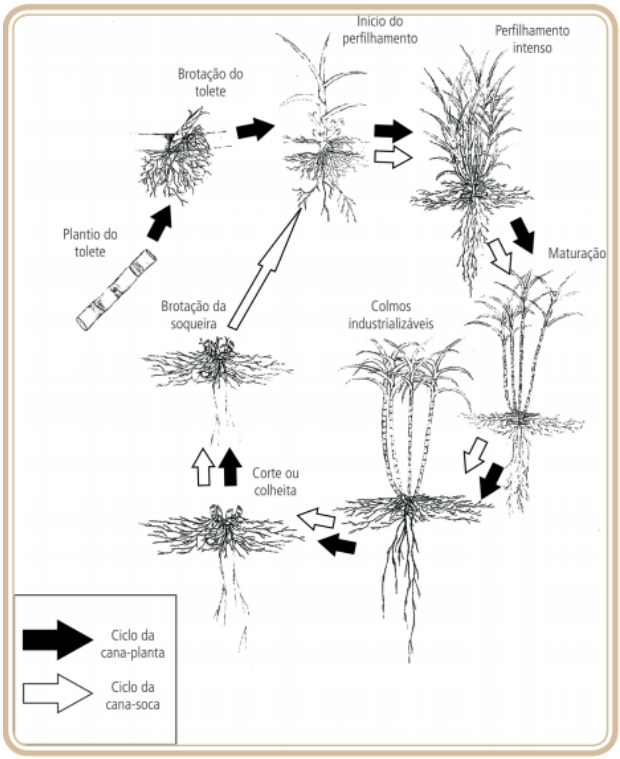


Figura 5 – Ciclo da cana-planta e da cana-soca.

Fonte: Segato et al. (2006)

O florescimento da cana-de-açúcar representa a capacidade de reprodução da planta, desejável principalmente em programas de melhoramento genético, mas que deve ser evitado em produções comerciais. Quando ocorre, o processo de formação do fruto há consumo de energia através da respiração de açúcares fotossintetizados, que, para o objetivo da produção, deveriam ser armazenados na forma de sacarose. Quando a energia é insuficiente para nutrição do fruto, pode haver consumo de sacarose que já estava armazenada nos entrenós. Este processo pode levar ao aumento de fibra e de bagaço e consequentemente ao menor valor de açúcar total recuperado (SEGATO et al., 2006).

Sendo suficiente a quantidade de água no solo, quando a semente é boa e não há ataque de pragas e doenças, é sempre satisfatória a porcentagem de brotação das gemas (CENTEC, 2004). O colmo é cortado em pedaços, chamados de toletes ou rebolos. O tolete contém dois a quatro nódios ou nós, conhecidos vulgarmente por gemas ou olhadura (SEGATO et al., 2006).

Antes de fazer o plantio, é importante conhecer a fertilidade do solo no qual se vai plantar a cana-de-açúcar. Para isso, é necessário fazer a análise do solo que irá identificar as necessidades e as quantidades de calcário e adubo mineral e/ou orgânico que devem ser

utilizados (CENTEC, 2004). Após o plantio, se houver condições ambientais favoráveis, principalmente de temperatura e umidade, irá iniciar o surgimento do broto na superfície da terra (RODRIGUES, 2017).

O surgimento das primeiras plantas varia com a profundidade do plantio, que depende da umidade do solo. Nos solos com bom teor de umidade, 4 a 5cm de cobertura será suficiente; solos frouxos e secos necessitam de um pouco mais de terra (CENTEC, 2004). Após 20 a 30 dias do plantio, vê-se os brotos em campo. Nesta fase as raízes e a parte aérea dos brotos estão iniciando seu crescimento, e a parte verde começa a adquirir sua capacidade fotossintética. Mesmo assim, ainda dependem das reservas nutritivas armazenadas no tolete para desenvolver (SEGATO et al., 2006).

É importante limpar o mato quando a cana está nascendo, ou ainda é nova, para evitar a matocompetição, pois roubam a água e nutrientes da terra, atrasando o crescimento e até provocando perdas significativas na produção de cana (CENTEC, 2004). Aproximadamente 20 a 30 dias após o surgimento dos primeiros brotos, observam-se novos brotos aparecendo. Como estas novas brotações (duas ou mais) tiveram sua origem do broto primário, denominam-se brotos secundários e dão início a fase do perfilhamento (KIMATI et al., 1997).

A fase de perfilhamento intenso da touceira se dá quando é atingido o máximo da produção de perfilhos. Quando começam a competir por luz, espaço, água e nutrientes, os mais jovens morrem e os mais velhos prosseguem seu crescimento. Ao final do perfilhamento, sem o dreno fisiológico dos colmos mais jovens, os colmos mais desenvolvidos continuam o seu crescimento em altura e espessura iniciando o processo de acúmulo de sacarose, como resultado da produção fotossintética (SEGATO et al., 2006).

No crescimento da parte aérea ocorre intensa divisão, diferenciação e alongamento celular, com grande aumento na massa seca da planta. No colmo ocorre a formação de nós, afastamento dos espaços entrenós e o desenvolvimento das folhas (AUDE, 1993). À medida que vão amadurecendo, ao atingir o seu tamanho final, passam a acumular mais intensamente a sacarose produzida pela fotossíntese constituindo os colmos industrializáveis (SEGATO et al., 2006).

A intensidade de acúmulo de sacarose é influenciada pelas condições ambientais que são desfavoráveis ao crescimento e desenvolvimento vegetativo, como temperaturas mais baixas, períodos de seca moderados e carência de nitrogênio. É nesta fase da cultura que se processa a colheita do canavial, anteriormente monitorado por análises específicas para a realização do corte (TORQUATO, 2006).

3.1.3 Fatores que influenciam o ciclo da cana-de-açúcar

Para os fatores ambientais, a temperatura e a umidade são variáveis críticas na brotação. Esta fase inicial exige temperaturas altas (30°C), enquanto temperaturas inferiores a 20°C e superiores a 35°C prejudicam a brotação da cana-de-açúcar, além disso, precisa

de boa umidade para que o processo ocorra de forma rápida. Excesso ou falta de água trarão problemas nesta fase do ciclo. Também doenças, como por exemplo a podridão abacaxi (fungo: *Ceratocystis paradoxa*) ou pragas, como a larva do besouro migdolus e cupins e, até mesmo plantas daninhas, como a tiririca (*Cyperus spp.*), podem resultar em falhas no canavial e consecutivas perdas. A textura e estrutura do solo estão diretamente relacionadas com a umidade e a geração do terreno (RODRIGUES, 2017) (SEGATO et al., 2006) (AUDE, 1993).

Sendo o clima favorável, a cana-de-açúcar pode ser cultivada em diversos tipos de solos, desde que haja umidade e os outros elementos com características parecidas. Deve ser, sobretudo, fértil, com boa drenagem e com bom teor de matéria orgânica. Os terrenos quando inclinados indicam a necessidade da realização de curvas de níveis, patamares ou terraços, que irão evitar uma provável erosão do solo. Esses corretivos no terreno conseguem segurar as águas das chuvas, manter as sementes, a matéria orgânica e adubos minerais no solo (CENTEC, 2004). Há significativa relação na produtividade da cana com a brotação (toletes ou soqueiras), o perfilhamento, o vigor das raízes e a boa brotação das socas (SEGATO et al., 2006). A Figura 6 ilustra os principais fatores que afetam a fase de brotação.

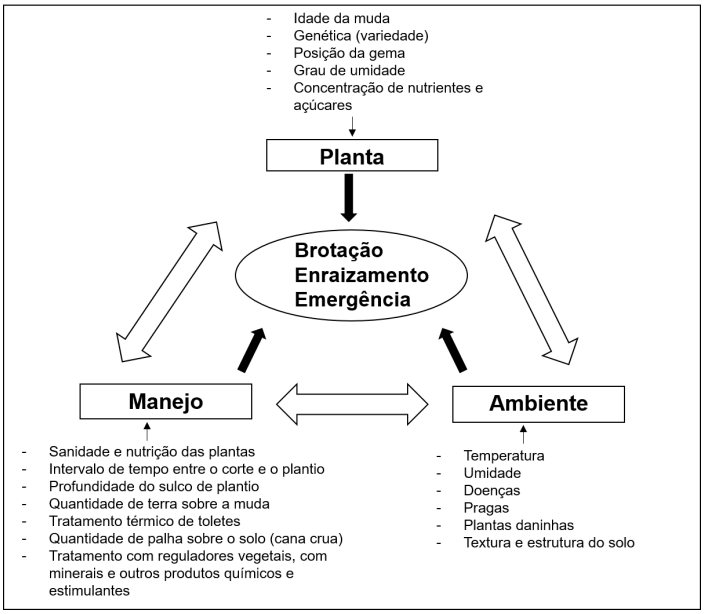


Figura 6 – Fatores que afetam a brotação, o enraizamento e a emergência da cana-planta.

Fonte: Adaptado de Segato et al. (2006)

A baixa luminosidade provocada pelo perfilhamento ou pelo excesso de plantas daninhas de grande porte reduz o surgimento de novos perfilhos. Os que sobrevivem à fase de competição terão seu crescimento acelerado, favorecendo a formação de colmos industrializáveis. Para o crescimento vegetativo, a disponibilidade de água, nutrientes e radiação solar são essenciais para o alongamento dos perfilhos, desenvolvimento das folhas

e crescimento dos colmos. Temperaturas do entre 25°C e 35°C favorecem o crescimento vegetativo da cultura (SEGATO et al., 2006) (AUDE, 1993).

Além da competição por luz, nutrientes e água, a matocompetição na cultura da cana-de-açúcar provoca redução drástica da colheita e elevação dos custos com controle e manejo da plantação. Ainda que sinta os efeitos do excesso hídrico e da deficiência hídrica, a fase de perfilhamento e crescimento dos colmos pode ser considerada como mais resistente a estes extremos do que a brotação (KIMATI et al., 1997).

Dependendo da velocidade do vento, da idade do canavial e do espaçamento utilizado, pode haver graves consequências durante o ciclo da cana. Poderá ocorrer faixa cloróticas nas folhas, o dilaceramento das lâminas foliares e o aumento da evapotranspiração. Já o acamamento[1], perde a energia armazenada como sacarose para nutrir brotos aéreos. Neste caso, a colheita do canavial fica mais difícil e custosa, pois além de perder matéria-prima no campo, leva muita impureza para a indústria (RODRIGUES, 2017).

Pragas, como a broca-da-cana-de-açúcar (*Diatraea saccharalis*) e a lagarta-elasmo (*Elasmopalpus lignosellus*), entre outro; assim como também os nematóides podem prejudicar o perfilhamento. Por outro lado, doenças como a ferrugem (fungo *Puccinia melanocephala*) e o carvão (fungo *Ustilago scitaminea*) tendem a aumentar o número de perfilhos (SEGATO et al., 2006). A Figura 7 ilustra os principais fatores que afetam a fase de perfilhamento.

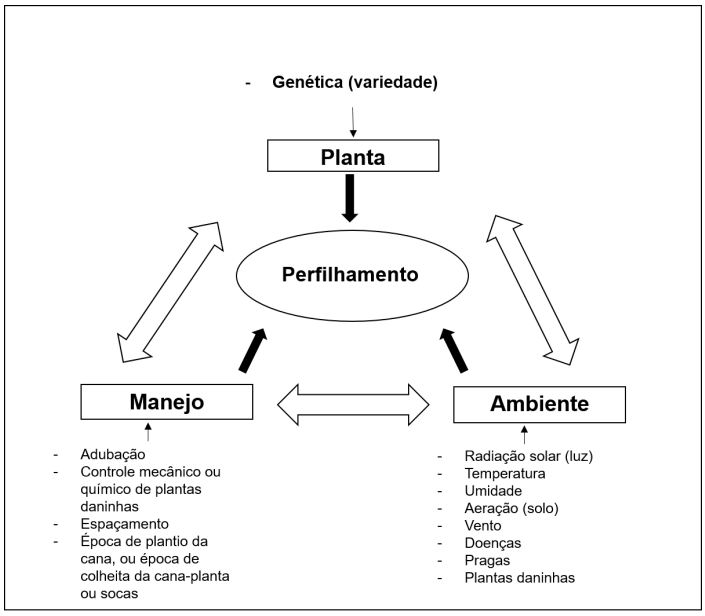


Figura 7 – Fatores que afetam o perfilhamento da cana-de-açúcar.

Fonte: Adaptado de Segato et al. (2006)

Pode-se definir a maturação da cana-de açúcar como o processo fisiológico de armazenamento de sacarose nos colmos. A maturação vai depender das características da planta,

[1] É a queda ou arqueamento das plantas em virtude da flexão do caule e/ou má ancoragem propiciada pelas raízes

mas as condições ambientais e de manejo podem favorecer ou não a maturação. Porém, sabe-se que as condições desfavoráveis ao crescimento vegetativo estimulam o armazanamento de sacarose (SEGATO et al., 2006).

A baixa temperatura é favorável ao acúmulo de sacarose, assim como a deficiência hídrica, inclusive o corte da irrigação, pode instigar a maturação. Como as folhas continuam realizando fotossíntese, a energia produzida é direcionada ao armazenamento de sacarose nos colmos. De modo contrário, o vento, quando muito forte, ao provocar acamamento no canavial, prejudica a maturação da cana-de-açúcar (SEGATO et al., 2006) (THOMAS, 2016). Portanto, para a maturação é recomendado temperaturas inferiores a 18-20ºC e/ou deficiência hídrica prolongada (CASAGRANDE, 1991).

Pragas e doenças podem afetar a maturação da cana. A broca dos colmos causa prejuízo direto (perda de tecido parenquimatoso), pela abertura de galerias. Os orifícios feitos pelas larvas, possibilitam a entrada de fungos, como o da podridão vermelha (*Colletotrichum falcatum*) que leva a inversão de sacarose previamente armazenada no colmo (KIMATI et al., 1997).

Solos com textura arenosa, porosos e mais secos favorecem a maturação, mas quando aplica-se muito nitrogênio no solo o crescimento vegetativo é estimulado enquanto a maturação é retardada. Logo, o ciclo da cana precisa de um período quente e úmido para brotar, emergir, perfilhar e outra parcialmente seco e/ou frio para o armazenamento da sacarose, e ter-se uma colheita rentável (RODRIGUES, 2017). A Figura 8 ilustra os principais fatores que afetam a fase de crescimento e maturação.

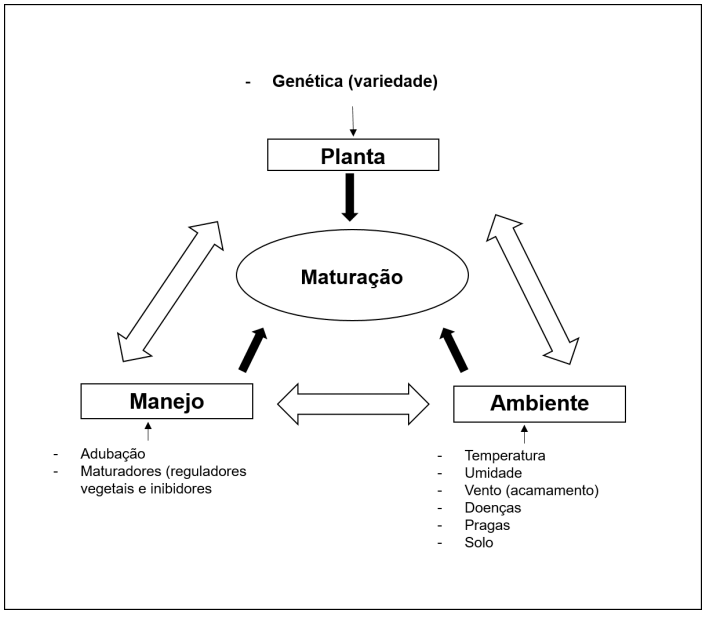


Figura 8 – Fatores que afetam a maturação (armazenamento de sacarose) da cana-de-açúcar.

Fonte: Adaptado de Segato et al. (2006)

3.1.4 Doenças e Pragas

Na cana-de-açúcar a maioria das doenças é controlada por causa do melhoramento genético que deixam a planta mais resistente a vários tipos de doenças. Entretanto, como a maioria das resistências a doenças na cultura da cana é quantitativa, a resistência é gradual, muitas variedades cultivadas podem apresentar níveis de suscetibilidade a doenças (SEGATO et al., 2006).

Segundo Segato et al. (2006), as 10 principais doenças que necessitam de maior controle são:

1. Doença causada por vírus: Mosaico;
2. Doenças causadas por bactérias: Escaldadura-das-folhas, Estrias vermelhas e Raquitismo-da-soqueira;
3. Doenças causadas por fungos: Carvão, Ferrugem, Mancha parda, Podridão abacaxi, Podridão de fusarium e Podridão vermelha.

3.1.5 Colheita

A verdadeira maturação da cana é a industrial, ou seja, quando o amido é transformado em açúcar e armazenado nas células parenquimatosas dos colmos. Este, não é igual ao da maturação fisiológica da planta, que é quando ela chega ao máximo do seu desenvolvimento apresentando folhas amareladas e mostrando que chegou ao fim do ciclo (CENTEC, 2004).

Na fase de amadurecimento não deve haver umidade excessiva no solo, caso contrário a cana irá brotar pelas gemas dos colmos e a sacarose servirá de fonte de energia para esses novos brotos. A falta de umidade do solo força a maturação, porque paralisa seu crescimento (KIMATI et al., 1997).

Quando os colmos industrializáveis estiverem, de acordo com as análises em campo e laboratório, aptos a serem colhidos, ou seja, em máxima maturação, ocorrerá a colheita da cana (TORQUATO, 2006). Para verificar se a cana está madura, na prática observa-se a folhagem amarelada, a casca rígida e brilhante e o colmo quebradiço (CENTEC, 2004).

Em campo, após o recolhimento da cana, os tocos conhecidos como rizomas que ficam do corte, brotam instantaneamente dando origem ao ciclo da cana-soca. Durante 20-30 dias aproximadamente, aos poucos o sistema radicular se renova principalmente em função da umidade do solo. Ocorre o perfilhamento, seguido da maturação e colheita dos colmos industrializáveis do novo ciclo, em número maior ou menor, dependendo da semente usada inicialmente e da interação da planta com o meio (SEGATO et al., 2006).

3.2 *Smart Agriculture* e Indústria 4.0 Aplicados na Agricultura

Como já foi discutido anteriormente, a agricultura possui um papel fundamental para a população mundial, sendo uma importante representante da economia de diversos países e necessária para suprir a demanda de alimentação e saúde do mundo. Dado a relevância da prática da agricultura em diferentes nações, há um grande incentivo em busca de novas tecnologias que possibilitem a otimização dos cultivos e manejos agropecuários, geralmente utilizando conceitos de IoT, MD, IA, *big data* e computação em nuvem.

O conceito de *Smart Agriculture*, ou agricultura de precisão, está normalmente associado à utilização de equipamentos de alta tecnologia (sejam *hardwares*, no sentido genérico do termo, ou *softwares*) para avaliar, ou monitorar, as condições numa determinada parcela de terreno, aplicando depois os diversos fatores de produção (sementes, fertilizantes, fitofármacos, reguladores de crescimento, água, dentre outros) (COELHO et al., 2004).

A evolução da informática, tecnologias em geoprocessamento, sistemas de posicionamento global e muitas outras tecnologias estão proporcionando à agricultura uma nova forma de se enxergar a propriedade agrária, podendo-se observar diversos fatores e características específicas. Esta mudança na forma de fazer agricultura está tornando cada vez mais o produtor rural um empresário rural, por controlar cada vez mais a linha de produção (TSCHIEDEL; FERREIRA, 2002).

A agricultura de precisão é uma filosofia de gerenciamento agrícola que parte de informações exatas, precisas e se completa com decisões exatas. Agricultura de precisão é uma maneira de gerir um campo produtivo metro a metro, levando em conta o fato de que cada pedaço da fazenda tem propriedades diferentes (ROZA, 2000).

Na maior parte dos casos, os principais fatores monitorados são: o tipo de solo, a capacidade de armazenamento de água, a umidade, o teor em nutrientes, o pH e a matéria orgânica. Outros fatores que também são muito importantes, porém mais difíceis de serem monitorados são: o declive, a exposição ao sol e a existência de pragas e/ou doenças, que são fatores igualmente responsáveis pela variabilidade espacial da produtividade das culturas (COELHO et al., 2004).

A agricultura de precisão envolve a medida dos fatores de produção, tendo em conta a variação espacial e temporal do potencial produtivo do meio e das necessidades específicas das culturas (RODRIGUES, 2017). Com este propósito, é popularmente dividido em dois grandes tipos de sistemas de monitoramento:

- O ambiental: Caracteriza a evolução de vários parâmetros do meio e das próprias plantas ao longo do tempo e no decurso da cultura;
- O da produtividade: Estima a variação espacial (no interior de uma parcela de cultura) da produção alcançada pela cultura.

O monitoramento da produtividade é a tecnologia de agricultura de precisão mais utilizada pelos agricultores dos países mais desenvolvidos, estando a sua aplicação muito difundida no caso das culturas de grãos. No entanto, há muito a se estudar e compreender deste universo complexo que envolve tantas dimensões e variáveis (COELHO et al., 2004).

Inovações nos campos da tecnologia da informação e da comunicação, do sensoriamento remoto, da instrumentação avançada, da automação e da robótica indicam que a agricultura de precisão emergirá como prática comum nas propriedades do futuro. Tais ferramentas e processos permitirão o uso da base de recursos naturais de forma mais inteligente, garantindo mais produtividade, eficiência e sustentabilidade à agricultura (LOPES; CONTINI, 2012). Todos esses conceitos da agricultura de precisão vêm relacionados a tecnologias utilizadas na indústria 4.0.

Com o desenvolvimento da internet, sensores cada vez menores e potentes, com preços cada vez mais acessíveis, *softwares* e *hardwares* cada vez mais sofisticados, a capacidade das máquinas aprenderem e colaborarem criando gigantescas redes (IoT), iniciou uma transformação na indústria, chamada de indústria 4.0, a qual consistem em uma visão de indústria conectada, sustentável e autônoma.

O impacto da Indústria 4.0 passa por uma forma muito mais complexa de inovação baseada na combinação de múltiplas tecnologias, que forçará as empresas a repensar a forma como gerem os seus negócios e processos, como se posicionam na cadeia de valor, como pensam no desenvolvimento de novos produtos e os introduzem no mercado, ajustando as ações de marketing e de distribuição (KIMATI et al., 1997). Olhando as grande monoculturas como indústrias, percebe-se que esses recursos da indústria 4.0 tendem a ser cada vez mais presentes nas plantações do futuro.

Sensores avançados viabilizarão o monitoramento de sistemas produtivos com grande precisão, novos materiais permitirão construir máquinas e equipamentos mais eficientes, precisos e duráveis, todos esses fatores serão fundamentais para o avanço e inovação da agricultura moderna, a criação de cultivos conectados e inteligentes (LOPES; CONTINI, 2012). Os computadores, tal como os conhecemos, tendem a desaparecer criando espaço para um novo conceito de computação ubíqua (COELHO, 2016).

O termo *big data* refere-se a grandes quantidades de dados, que são armazenados a cada instante, resultante da existência de milhões de sistemas atualmente ligados à rede (IoT), produzindo dados em tempo real (COSTA, 2017). Com tantos dados a serem gerados continuamente são precisas ferramentas de análise poderosas para lhes dar significado. Dados são números, palavras ou outros sinais e representam fatos discretos sobre uma realidade objetiva. Podem ser verificados e validados, contudo não tem qualquer significado se não forem interpretados e contextualizados, dando origem à informação. A informação tende a evoluir levando à criação de teorias e a previsões de futuro, neste caso tem-se o conhecimento (TORQUATO, 2006)(COELHO, 2016).

Nota-se que a indústria 4.0 e a agricultura de precisão estão fortemente focadas na

melhoria contínua em termos de eficiência, segurança, qualidade e produtividade das produções. Porém, o grande desafio é colecionar todos os dados considerados relevantes, processá-los, transformando-os em conhecimento. Esta atividade requer sistemas tecnologicamente evoluídos, providos de capacidade de processamento em tempo próximo ao real e algoritmos sofisticados. Alcançar o conhecimento e a sabedoria abre horizontes para além do imaginário, sendo um grande motor para o mundo e do caminho para a indústria do futuro (COELHO, 2016).

3.3 Mineração de Dados

Mineração de Dados (MD), ou *Data Mining*, é um ramo da computação que teve início nos anos 80, quando os profissionais das empresas e organizações começaram a se preocupar com os grandes volumes de dados digitais estocados e inutilizados dentro da empresa. Nesta época, MD consistia essencialmente em extrair informação de gigantescas bases de dados da maneira mais automatizada possível. Atualmente, MD consiste, sobretudo, na análise dos dados após a extração buscando-se, por exemplo, levantar as necessidades reais e hipotéticas de cada indivíduo de uma cidade. Amo (2004) define os seguintes pontos como algumas das razões do porquê da MD se tornar necessária para práticas de gestão e tomadas de decisão: (a) os volumes de dados são muito importantes para um tratamento utilizando somente técnicas clássicas de análise, como cálculos manuais, (b) o usuário final não é necessariamente um estatístico e (c) a intensificação do tráfego de dados na *web 2.0* (navegação na *internet*, *e-commerce*, redes sociais, catálogos *online*, etc.) aumenta a possibilidade de acesso aos dados.

O foco central da MD é o de como transformar dados armazenados em conhecimento, expresso em termos de formalismos de representação, tal como regras e relações entre dados. Existe conhecimento que pode ser extraído diretamente de dados sem o uso de qualquer técnica. Entretanto, existe também muito conhecimento que está de certa forma "embutido" na base de dados, na forma de relações existentes entre itens de dados que, para ser extraído, e necessário o desenvolvimento de técnicas especiais (REZENDE et al., 2003).

Na caracterização da MD é importante discutir o que na literatura é chamado de *Knowledge Discovery in Databases* (KDD), ou seja, Descoberta de Conhecimento em Bases de Dados. De acordo com Frawley, Piatetsky-Shapiro e Matheus (1992), KDD é a extração de conhecimento previamente desconhecido, implícito e potencialmente útil, a partir de dados. Para Fayyad, Piatetsky-Shapiro e Smyth (1996) é o processo que busca descobrir correlacionamentos e dados implícitos nos registros de um banco de dados, estudando e desenvolvendo um processo de extração de conhecimento novo, útil e interessante e apresentá-lo de alguma forma acessível para o usuário.

O termo KDD, foi formalizado em 1989 em referência ao amplo conceito de procu-

rar conhecimento a partir de base de dados. O KDD é um processo, de várias etapas, não trivial, interativo e iterativo, com o objetivo de identificar padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. O termo iterativo sugere a possibilidade de repetições integrais ou parciais do processo de KDD, e a expressão não trivial alerta para a complexidade normalmente presente na execução deste processo. Já com relação à expressão padrão válido indica que o conhecimento deve ser verdadeiro e adequado ao contexto da aplicação, e o termo padrão novo deve acrescentar novos conhecimentos aos existentes, para que todo esse processo gere conhecimento útil que possa ser aplicado de forma a proporcionar benefícios ao contexto de aplicação de KDD.

Porém, a extração de conhecimento de uma grande base de dados através da aplicação de um processo de KDD exige a melhor compreensão das diferenças entre dado (códigos que constituem a matéria prima da informação), informação (dados tratados que possuem significado) e conhecimento (que além do significado possui uma aplicação). O processo de transformação de uma base de dados até a entrega de conhecimento para se poder gerar inteligência aumenta proporcionalmente ao nível de independência do contexto, ou seja, é uma informação refinada a partir dos dados e do entendimento que pode ser compreendida pelos usuários (TEOFILO, 2015), conforme ilustrado na Figura 9.

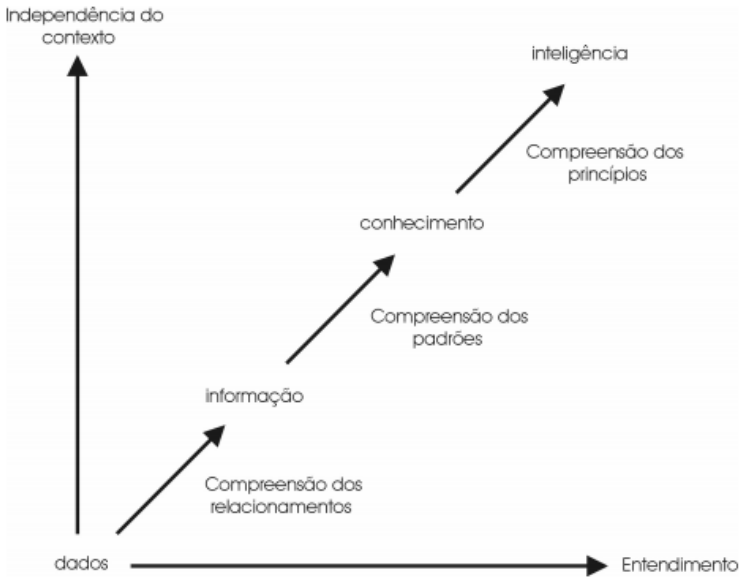


Figura 9 – Dado, informação e conhecimento

Fonte: Kock Jr, McQueen e Baker (1996)

O KDD é dividido em cinco etapas, como pode ser visto na Figura 10: seleção, pré-processamento, transformação, mineração de dados e análise dos resultados. O objetivo desse processo é descobrir informações relevantes e importantes para apoiar os tomadores de decisão em suas decisões estratégicas.

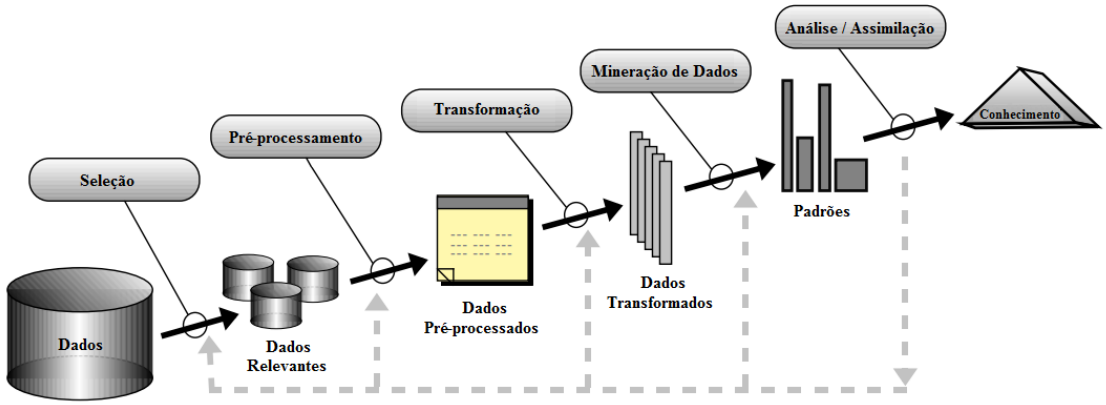


Figura 10 – Etapas do processo KDD

Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996)

Segundo Teofilo (2015) pode-se descrever melhor as cinco etapas do KDD em sete tópicos:

1. Limpeza dos dados: etapa onde são eliminados ruídos e dados inconsistentes.
2. Integração dos dados: etapa onde diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados.
3. Seleção: etapa onde são selecionados os atributos que interessam ao usuário. Por exemplo, o usuário pode decidir que informações como nome e telefone não são de relevantes para decidir se um cliente é um bom comprador ou não.
4. Transformação dos dados: etapa onde os dados são transformados num formato apropriado para aplicação de algoritmos de mineração (por exemplo, através de operações de agregação).
5. Mineração: etapa essencial do processo, consistindo na aplicação de técnicas de AM a fim de se extrair os padrões de interesse.
6. Avaliação ou Pós-processamento: etapa onde são identificados os padrões interessantes de acordo com algum critério do usuário.
7. Visualização dos Resultados: etapa onde são utilizadas técnicas de representação de conhecimento a fim de apresentar ao usuário o conhecimento minerado.

Outra metodologia muito utilizada no processo de mineração de dados é a *CRoss Industry Standard Process for Data Mining* (CRISP-DM). O CRISP-DM é uma das metodologias mais populares para aumentar o sucesso dos processos de MD (CHAPMAN

et al., 2000). A metodologia define uma sequência não rígida de seis fases, que permite a construção e implementação de um modelo de mineração para ser usado em um ambiente real, ajudando as decisões de negócios (MORO; LAUREANO; CORTEZ, 2011).

No desenvolvimento desse trabalho utilizou-se as fases do CRISP-DM nos processos de MD necessários. Essas fases são ilustradas na Figura 11 e descritas em sequência.

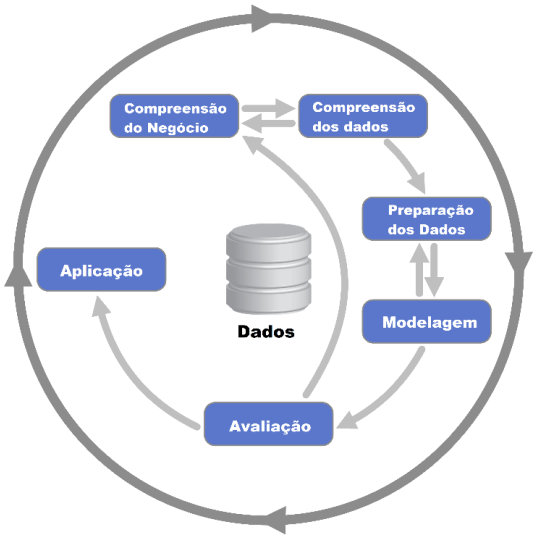


Figura 11 – Fases do CRISP-DM

Fonte: Adaptado de Chapman et al. (2000)

1. Compreensão do Problema: Essa fase inicial se concentra na compreensão dos objetivos e requisitos do problema abordado, convertendo esse conhecimento em uma definição de problema de MD.

2. Compreensão do Negócio (Problema): A fase de compreensão dos dados permite familiarizar-se com os dados, identificar problemas de qualidade, descobrir primeiros *insights* sobre os dados e detectar subconjuntos interessantes para formar hipóteses sobre informações relevantes.

3. Preparação dos Dados: Nessa etapa, as tarefas incluem seleção de atributos, além de transformação e limpeza de dados para melhor aplicação das técnicas.

4. Modelagem: Nessa fase, técnicas de MD são selecionadas e aplicadas. Representa o desenvolvimento dos modelos para o problema, com base nos dados que já foram adequados para serem utilizados.

5. Avaliação: Nessa fase do projeto o modelo desenvolvido é avaliado para ratificar a sua adequação ao problema em estudo.

6. Aplicação: Todo conhecimento obtido por meio do trabalho de mineração tornam-se subsídios para o desenvolvimento de estratégias para tomada de decisão no contexto do problema estudado.

Com o uso da MD, é possível descobrir informações relacionadas a associações, sequências, classificação, aglomeração (*clustering*) e prognósticos. Esses sistemas realizam uma análise de alto nível quanto a padrões ou tendências, mas também podem esmiuçar os dados para revelar mais detalhes, se necessário. A próxima subseção apresenta as principais técnicas utlizadas nesta pesquisa durante as etapas de mineração para a obtenção de informação.

3.3.1 Análise de Dados e Técnicas Utilizadas na Mineração

A análise das bases de dados obtidas é fundamental para a obtenção das informações desejadas, essas análises envolvem desde métricas estatísticas tradicionais, de medidas de posição e dispersão, a outros métodos mais sofisticados para a obtenção de informações.

Análise de dados é o processo de transformar um conjunto de dados a fim de poder verificá-los melhor, dando-lhes, uma razão de ser e uma análise racional. Trata-se de analisar os dados de um problema e identificá-los. A análise de dados possui diferentes abordagens, incorporando técnicas diversas (HAIR et al., 2009).

Basicamente, os dados podem ser classificados como quantitativos ou qualitativos (SHEATS; PANKRATZ, 2002). Dados quantitativos são geralmente medidos em uma escala contínua. Após a coleta dos dados, é possível obter uma medida de tendência central e um indicador da variabilidade dos dados. A medida de tendência central mais usada para os dados numéricos é a média (a Equação 1 ilustra média aritmética e a 2 a ponderada), enquanto o desvio-padrão é o estimador de variabilidade mais comumente empregado quando a variável examinada é do tipo contínua ou paramétrica (NORMANDO; TJÄDERHANE; QUINTÃO, 2010).

$$M_a = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$M_p = \frac{\sum X_i.f_i}{\sum f_i} \tag{2}$$

Uma variável qualitativa ou categórica apresenta um número limitado de valores ou categorias, e pode ser classificada em ordinal ou nominal. Dados ordinais devem seguir um nível crescente (ordem) entre as categorias, a medida de tendência central mais comumente usada para dados ordinais é a mediana, que é o ponto médio a partir do qual metade dos valores coletados é superior a esse ponto e a outra metade é inferior. Por outro lado, os dados nominais são distribuídos em categorias, onde nenhuma ordem inerente pode ser observada. Esse tipo de dado formado por números inteiros (variável discreta) é, rotineiramente, descrito como frequência absoluta ou relativa (percentagem) (NORMANDO; TJÄDERHANE; QUINTÃO, 2010).

Ainda é importante ressaltar que com dados quantitativos pode-se aplicar testes paramétricos, exigindo que sejam cumpridos três pressupostos: distribuição normal, homo-

geneidade dos dados e variáveis intervalares e contínuas. Já com os qualitativos pode-se realizar os testes não paramétricos, que são muito úteis para a análise de testes de hipóteses (GONÇALVES; MARCONDES; LAKATOS, 1982). São também úteis para a análise de amostras grandes, em que os pressupostos paramétricos não se verifiquem, assim como para as amostras muito pequenas e para as investigações que envolvam hipóteses cujos processos de medida sejam ordinais. Além disso, os testes não paramétricos não são tão fidedignos como os testes paramétricos, ou seja, com os testes não paramétricos não se encontram tantas diferenças entre os dados, quando estas diferenças realmente existem (MARÔCO, 2011). Diversas análises são realizadas durante este trabalho para a obtenção de informações das bases de dados.

As medidas de posição, ou de tendência central, fornecem medidas que podem caracterizar o comportamento dos elementos de uma série, possibilitando determinar se um valor está entre o maior e menor valor da amostra, ou se está localizado no centro do conjunto de dados, por exemplo. São indicadores que permitem que se tenha uma primeira ideia, ou um resumo, do modo como se distribuem os dados de uma experiência informando sobre o valor da variável aleatória (MORAIS, 2005). As medidas de posição mais importantes são os diferentes tipos de média, a mediana e a moda.

Uma medida de posição interessante são os quartis. Eles dividem a distribuição em quatro partes iguais de 25%, tem-se assim 3 quartis numa distribuição. O 1º quartil $Q1$ (Equação 3) separa os 25% de dados inferiores, o 2º quartil $Q2$ (Equação 4) separa os 50% de dados inferiores e o 3º quartil $Q3$ (Equação 5) separa os 75% de dados inferiores. Os Quartis também podem ser entendidos como os percentis: $P25$, $P50$, $P75$ e observa-se que $Q2$ é igual a mediana da amostra (ARALDI, 2005). A obtenção dos quartis é interessante pois através deles pode-se identificar medidas de dispersão, como o desvio quartílico, útil para a obtenção de *outliers*.

$$Q_1 = [p(sup) - 0,25].x(inf) + [0,25 - p(inf)].x\frac{(sup)}{p(sup)} - p(inf) \tag{3}$$

$$Q_2 = [p(sup) - 0,50].x(inf) + [0,50 - p(inf)].x\frac{(sup)}{p(sup)} - p(inf) \tag{4}$$

$$Q_3 = [p(sup) - 0,75].x(inf) + [0,75 - p(inf)].x\frac{(sup)}{p(sup)} - p(inf) \tag{5}$$

As medidas de dispersão traduzem a variação de um conjunto de dados em torno da média ou mediana, por exemplo, ou seja, da maior ou menor variabilidade dos resultados obtidos. Permitem identificar até que ponto os resultados se concentram ou não ao redor da tendência central de um conjunto de observações. Incluem, entre outras, o desvio absoluto médio (Equação 6), a variância (Equação 7), o desvio padrão (Equação 8), o coeficiente de variação (Equação 9) e o desvio quartílico (Equação 10), cada uma expressando diferentes formas de quantificar a tendência que os resultados de uma expe-

riência aleatória têm para se concentrarem em determinados valores. Quanto maior for a dispersão, menor é a concentração, e vice-versa (MORAIS, 2005). A variabilidade (ou dispersão) de um conjunto de dados pode ser quantificada através da amplitude de variação, da variância, do desvio-padrão do coeficiente de variação, entre outras (BASTOS; DUQUIA, 2007).

$$D_M = \frac{1}{n}\sum_{i=1}^{n} |xi - media| \tag{6}$$

$$S^2 = \frac{\sum(X - media)^2}{(n-1)} \tag{7}$$

$$S = \sqrt{\frac{\sum(X - media)^2}{(n-1)}} \tag{8}$$

$$C_v = \frac{S}{media} \tag{9}$$

$$DQ = \frac{Q_3 - Q_1}{2} \tag{10}$$

A partir das análises das medidas de posição e dispersão de uma amostra de dados pode-se obter muitas informações relevantes e importantes para a tomada de decisão, porém, outras métricas devem ser analisadas para a obtenção de informações mais precisas de uma base de dados. A identificação de correlações e *outliers* são importantes caracterizadores para as amostras. Através destas análises pode-se realizar técnicas de séries temporais e regressores para a prospecção de comportamentos futuros dos dados estudados.

Em teoria da probabilidade e estatística, a correlação indica a força e a direção do relacionamento linear, ou não linear, entre duas variáveis aleatórias. No uso estatístico geral, a correlação se refere a medida da relação entre duas variáveis, embora correlação não implique causalidade.

Muitas vezes é preciso conhecer a forma como duas ou mais variáveis estão relacionadas. Existem diversos critérios de avaliação dessa relação, alguns próprios para as que seguem uma distribuição normal e outros para as que não seguem uma distribuição teórica conhecida. Basicamente, existem métodos de avaliação da relação para variáveis contínuas e categóricas (discretas) (GUIMARÃES, 2008). Neste sentido geral, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados.

Para Moore (2007), a correlação mensura a direção e o grau da relação linear entre duas variáveis quantitativas. O coeficiente de correlação de Pearson (r) é uma medida

de associação linear entre variáveis (FIGUEIREDO-FILHO; SILVA-JUNIOR, 2010). Sua fórmula é ilustrada na Equação 11.

$$r = \frac{\sum_{i=1}^{n}(x_i - x')(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2}.\sqrt{\sum_{i=1}^{n}(y_i - y')^2}} \tag{11}$$

Onde x' (Equação 12) e y' (Equação 13) são as médias aritméticas.

$$x' = \frac{1}{n}.\sum_{i=1}^{n} x_i \tag{12}$$

$$y' = \frac{1}{n}.\sum_{i=1}^{n} y_i \tag{13}$$

O coeficiente de correlação Pearson (r) varia de -1 a 1. O sinal indica direção positiva ou negativa do relacionamento e o valor sugere a força da relação entre as variáveis. Uma correlação perfeita (-1 ou 1) indica que o escore de uma variável pode ser determinado exatamente ao se saber o escore da outra. No outro oposto, uma correlação de valor zero indica que não há relação linear entre as variáveis, não existindo um determinante para a correlação (FIGUEIREDO-FILHO; SILVA-JUNIOR, 2010).

Para Cohen (1988), pontuações entre 0,10 e 0,29 podem ser consideradas pequenas; índices entre 0,30 e 0,49 podem ser considerados como médios; e valores entre 0,50 e 1 podem ser interpretados como grandes. Dancey e Reidy (2013) apontam para uma classificação ligeiramente diferente: $r = 0{,}10$ até 0,30 (fraco); $r = 0{,}40$ até 0,6 (moderado); $r = 0{,}70$ até 1 (forte). Seja como for, o certo é que quanto mais perto de 1, ou -1, (independente do sinal), for o coeficiente, maior é o grau de dependência estatística linear entre as variáveis. No outro oposto, quanto mais próximo de zero, menor é a força dessa relação.

Para estimar a correlação de dois conjuntos que não têm distribuição conjunta normal bivariada, a alternativa mais usual é o coeficiente de correlação de Spearman. A correlação de Spearman (ou rho) é um cálculo estatístico baseado em postos e foi introduzido por Spearman (1904) que exige apenas que as variáveis X e Y sejam medidas pelo menos em escala ordinal. Este coeficiente é uma medida de correlação não-paramétrica, isto é, ele avalia uma função monótona arbitrária que pode ser a descrição da relação entre duas variáveis, sem fazer nenhumas suposições sobre a distribuição de frequência. Ao contrário do coeficiente de correlação de Pearson, o coeficiente de Spearman não requer a suposição de que a relação entre as variáveis é linear, nem que elas sejam medidas em intervalo de classe, podendo ser usado para os conjuntos medidos no nível ordinal.

De acordo com Siegel e Castellan Jr (1975) o coeficiente de correlação por postos de Spearman, designado e representado por r_s ou rho, é uma medida não paramétrica que permite estabelecer a existência de correlação entre duas variáveis. Para o cálculo de r_s considera-se um conjunto de n pares de observações. Em correspondência a cada par, consta-se seu posto em relação uma determinada variável I_j, o índice ambiental e seu posto

em relação à outra variável, neste caso, Y_ij, que em geral é representada pela produção do i-ésimo genótipo no j-ésimo ambiente. Em seguida, determinam-se os valores da diferença, denotada por d_i, entre os pares de postos. Eleva-se cada valor de d_i ao quadrado e soma-se, obtendo-se assim a soma de quadrados da diferença entre os pares de postos, como pode ser visto na Equação 14.

$$\sum_{i=1}^{n} d_i^2 \tag{14}$$

O coeficiente de correlação de Spearman é obtido por meio da Equação 15.

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{15}$$

Para amostras superiores a dez (10) elementos, segundo Siegel e Castellan Jr (1975), a significância de um valor r_s obtido pode ser verificada através da identificação de uma aproximação para a distribuição t de Student com a Equação 16, que é uma distribuição de probabilidade teórica, simétrica, campaniforme e semelhante à curva normal (YUE; PILON, 2004).

$$t = \frac{r_s}{\sqrt{\frac{(1-r_s^2)}{(n-2)}}} \tag{16}$$

Segundo Figueiredo Filho e Silva Junior (2010) um dos objetivos centrais da estatística é fazer inferências válidas para a população a partir de dados amostrais. O valor p (*p-value*) apresenta a probabilidade dos valores encontrados a partir de dados amostrais serem representativos dos parâmetros populacionais, dado que a hipótese nula é verdadeira. Quanto menor for o *p-value*, maior é a confiança do pesquisador em rejeitar a hipótese nula. No outro oposto, valores altos do valor p indicam que a hipótese nula não pode ser rejeitada. Para Moore (2007), a probabilidade estimada, assumindo que Ho é verdadeira, de que a estatística assumiria um valor extremo ou maior do que foi de fato observado, é chamado de valor p. Fazendo uma generalização pode-se dizer que para um $p - value = 0,1$, há um nível significativo de 10% de chance do pesquisador estar errado diante da hipótese nula.

Segundo Sigel e Castellan Jr (1975), Bauer (2007) e Pontes (2010), em se tratando de correlação de Spearman deve-se atentar para o tamanho da amostra analisada. Mesmo o valor mínimo apontado sendo a partir de dez dados amostrais, a correlação pode seguir uma distribuição não linear e o valor p pode não ser correto devido à expansão artificial desta amostra.

Em processos de investigação de amostra de dados pode-se deparar com conjuntos em que algumas observações se afastam demasiadamente das restantes, parecendo que foram geradas por um mecanismo diferente. O estudo destas observações é importante dado

que uma das importantes etapas, em qualquer análise estatística de dados, é estudar a qualidade das observações (MUNOZ-GARCIA J., 1990).

As observações que apresentam um grande afastamento das restantes, ou são inconsistentes com elas, são habitualmente designadas por *outliers*. Pode-se concluir que um *outlier* é caracterizado pela sua relação com as demais observações que fazem parte da amostra. O seu distanciamento em relação a essas observações é fundamental para se fazer a sua caracterização. Estas observações são também designadas por observações "anormais", contaminantes, estranhas ou aberrantes (FIGUEIRA, 1998).

Na análise de séries temporais, encontram-se frequentemente *outliers* e mudanças estruturais, que podem estar associadas a acontecimentos inesperados ou incontroláveis, como por exemplo, guerras, mudanças políticas, ou podem simplesmente ocorrer devido a erros de medição ou de registo de observações. Estas observações podem comprometer os procedimentos usuais de modelação linear de uma série temporal, nomeadamente podem induzir a uma identificação incorreta de um modelo ARIMA - que é um modelo muito utilizado na modelagem e previsões de séries temporais - e a uma estimação enviesada dos parâmetros do modelo (MIRANDA, 2001).

A distribuição normal dos valores medidos determina a faixa de erros aleatórios. Os erros de medida fora desta faixa são denominados dispersos. Existem vários testes de rejeição de dados e, dependendo do teste, em um mesmo conjunto podem ser detectados um ou mais resultados suspeitos. Isso é mais do que suficiente para provar que tais testes não podem ser encarados como disciplina rotineira (OLIVEIRA, 2008).

Como segue, existem cáculos para a obtenção de *outliers*. Considerando que a amostra esteja ordenada e que se tenha os valores Q1 (Primeiro quartil) e Q3 (Terceiro quartil), deve-se calcular o intervalo inter-quartil, ou *Interquartile Range* (IQR), como é demonstrado na equação 17, e com ele determinar o *Lower Outlier Boundary* (LOB), Equação 18, e o *Upper Outlier Boundary* (UOB), Equação 19.

$$(IQR) = Q3 - Q1 \tag{17}$$

$$LOB = Q1 - (1.5 * IQR) \tag{18}$$

$$UOB = Q3 + (1.5 * IQR) \tag{19}$$

A operação dos cálculos de LOB e de UOB apontam as barreias de investigação dos *outliers*, ou seja, qualquer valor que for identificado na amostra que não esteja entre os valores encontrados pode ser apontado como um dado discrepante. Quando o fator de multiplicação utilizado é o 1,5 se encontram as barreiras internas de investigação, caso se deseje aumentar a aceitação a erros pode-se multiplicar a amplitude interquartílica por 3 (MCGILL; TUKEY; LARSEN, 1978), determinando um intervalo onde qualquer valor identificado fora dele é considerado um *outlier* extremo.

O *boxplot* (gráfico de caixa) é uma ferramenta importante para ser utilizada na observação de amostras em conjunto com a identificação de *outliers*. Ele permite efetuar uma análise comparativa dos dados, numa tentativa de determinar a sua tendência central, a sua dispersão, a sua assimetria e a existência de *outliers*. O *boxplot* ilustra uma série de valores e possibilita distintas análises como pode ser visto na Figura 12 (WILLIAMSON; PARKER; KENDRICK, 1989). O Q2 centralizado na caixa indica uma amostra simétrica, Q2 próximo ao Q1, indica uma assimetria positiva, já Q2 próximo ao Q3, indica uma assimetria negativa.

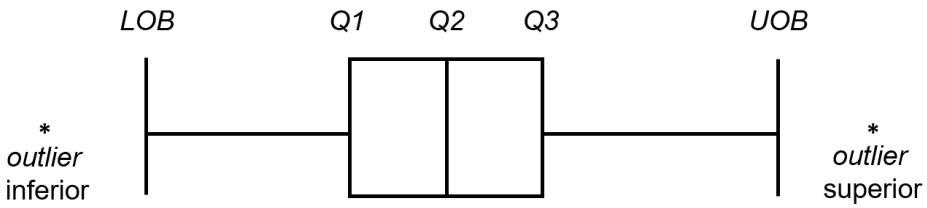


Figura 12 – Informações em um *box-plot*

Fonte: Adaptado de Williamson, Parker e Kendrick (1989)

A MD realizada no estudo de caso desta pesquisa utilizou todas as medidas de posição, dispersão e técnicas descritas nessa seção. A obtenção de medidas de posição, diante das análises dos dados, foram aplicadas para viabilizar uma inferência generalista e de distribuição dos dados. A identificação de *outliers* é útil para a validação das amostras, identificando dados anormais para serem excluídos das análises. A aplicação de técnicas de correlação possibilita o cruzamento de bases de dados e métricas até então não relacionadas, buscando um indicativo de correlação significante.

3.4 Sistemas Complexos

Um sistema complexo é um conjunto de partes, em diferentes escalas e níveis de organização, interligadas formando uma estrutura estável (HEYLIGHEN, 1988). Os sistemas complexos apresentam comportamentos emergentes, que se manifestam como resultado das interações entre as partes. Assim, a totalidade do sistema é maior que a soma de suas partes e o todo exibe padrões e estruturas que surgem espontaneamente de suas partes. As partes são distintas mas interconectadas, não sendo possível analisar um sistema complexo por métodos reducionistas (LOZANO, 2017).

Para a modelagem de um sistema complexo normalmente considera-se previsibilidade espacial e temporal, não sendo necessário conhecer todo o sistema para reconstruí-lo ou prever sua estrutura, o sistema é redundante. Assim, a complexidade é considerada particularmente por meio do conceito complexidade estrutural que é a quantidade de informação que o sistema armazena. A evolução de sistemas complexos é marcada por um

misto de ordem (muitas conexões, poucas ou nenhuma distinção no padrão comportamental) e desordem (muitas distinções no padrão comportamental e poucas conexões). É fato que algumas partes ou estruturas do sistema serão conservadas durante uma certa evolução de tempo enquanto que outras irão se modificar (LOZANO, 2017).

Como exemplo de um sistema complexo pode-se pensar em uma plantação e em todas as variáveis e fatores que interferem no seu cultivo. Analisando apenas o tipo de plantação, ou a temperatura, ou a velocidade dos ventos, individualmente, não se tem uma visão completa do estado de cultivo desta plantação, nem muito menos pode-se fazer análises para tomadas de decisão sobre esta. Fenômenos ambientes são representantes de sistemas complexos por apresentarem características inconstantes e que dependem de um conjunto de variáveis para poderem ser estimadas.

Um sistema complexo segue, em suma, quatro propriedades (BAR-YAM, 2003):

- ❏ Unidade Coletiva: um Sistema Complexo é composto por um conjunto de partes conectadas por alguma forma de inter-relação entre elas. Assim, para caracterizar um sistema é necessário não somente conhecer as partes, mas também os modos de relação entre elas.

- ❏ Organicidade funcional: em um Sistema Complexo cada subsistema possui um processamento interno de informações, de modo que ocorre uma relação funcional entre os subsistemas.

- ❏ Propriedade emergente: as interações entre as partes de um Sistema Complexo criam um padrão coletivo chamado propriedade emergente. Estas propriedades consistem uma exteriorização do Sistema Complexo. Em outras palavras, a dinâmica das partes em uma escala de relação produz uma propriedade emergente em um nível mais alto de escala.

- ❏ Multiescalas: no estudo dos Sistemas Complexos ocorrem sistemas interagindo com outros sistemas, de modo a formar Sistemas mais amplos em escalas e com propriedades emergentes.

O estudo das propriedades emergentes nos sistemas complexos possibilita a compreensão de seu comportamento global e da forma pela qual os componentes individuais se interagem para formar o todo. Entender o comportamento de sistemas complexos envolve, portanto, a compreensão do padrão típico de estruturação, das propriedades similares que independem da escala de observação, e da dinâmica temporal predominantemente não-linear, sensível a variações externas e cuja previsibilidade de longo alcance é mais improvável (SOUZA; BUCKERIDGE, 2004). Dessa forma, no estudo dos sistemas complexos, é importante considerar dois aspectos:

- ❏ Como as partes do sistema desenvolvem computações e lidam com mudanças aleatórias;
- ❏ Como a informação é integrada e como isso dispara a emergência de uma computação de estados.

Dentre as técnicas utilizadas para modelagem de sistemas complexos tem-se equações diferenciais, cadeias de Markov, AC, teoria de agentes e redes complexas. Apesar de todos os avanços computacionais, o que ainda não se compreende é muito vasto (LOZANO, 2017).

3.5 Cadeias de Markov

Ao se observar e analisar fenômenos reais, percebe-se que a complexidade e a incerteza os acompanham, principalmente quando o fenômeno observado envolve dinâmicas da natureza e do comportamento humano e ações imprevisíveis. Os modelos determinísticos certamente contribuem para a compreensão, a um nível básico, do comportamento dinâmico de um sistema. No entanto, por não poderem lidar com a incerteza, acabam por ser insuficientes nos processos de tomada de decisão. Assim, recorre-se a processos estocásticos como uma forma de tratar quantitativamente estes fenômenos, aproveitando certas características de regularidade que eles apresentam para serem descritos por modelos probabilísticos (ALVES; DELGADO, 1997).

Pode definir-se um processo estocástico como um conjunto de variáveis aleatórias indexadas a uma variável (geralmente a variável tempo), sendo representado por $X(t)$. Estabelecendo o paralelismo com o caso determinístico, onde uma função $f(t)$ toma valores bem definidos ao longo do tempo, um processo estocástico toma valores aleatórios ao longo do tempo. Aos valores que $X(t)$ pode assumir, chamam-se estados e ao seu conjunto X, espaço de estados (ROSS, 1983).

Muitos fenômenos que ocorrem na natureza e na sociedade podem ser estudados, pelo menos em uma primeira observação, como se os fenômenos começassem a partir de um estado inicial e passassem por uma sequência de estados, em que a transição de um estado para o seguinte, ocorre segundo uma certa probabilidade. No caso em que esta probabilidade de transição depende apenas do estado em que o fenômeno se encontra e do estado a seguir, o processo é denominado de Processo de Markov e uma sequência de estados envolvida nesse processo é denominada de cadeias de Markov (BOLDRINI et al., 1980).

Um processo de Markov é um processo estocástico em que a probabilidade do sistema estar no estado i no período $(n + 1)$ depende somente do estado em que o sistema está no período n. Ou seja, para os processos de Markov, só interessa o estado imediato. Os

principais elementos de um processo de Markov são dois (SIMON; BLUME; DOERING, 2004):

- a probabilidade $x^i(n)$ de ocorrer o estado i no n-ésimo período de tempo, ou, alternativamente, a fração da população em questão que está no estado i no n-ésimo período de tempo;
- as probabilidades de transição m_{ij}, que representam as probabilidades de o processo estar no estado i no tempo $(n+1)$ dado que está no estado j no tempo n. Estas probabilidades de transição são normalmente agrupadas numa matriz, denominada de matriz de transição, matriz estocástica ou ainda matriz de Markov.

Para uma definição de cadeia de Markov, pode-se considerar que as mudanças de estado em determinados instantes de tempo n sejam denotadas por X_n e que pertençam a um conjunto S de estados possíveis, chamado espaço de estado. Supondo que $S = 1, ..., M$, para algum inteiro positivo m. As cadeias de Markov são descritas em termos das probabilidades de transição p_{ij}, como visto na Equação 20.

$$p_{ij} = P(X_{n+1} = j | X_n = i), i, j \epsilon S \tag{20}$$

O pressuposto fundamental dos processos de Markov é que as probabilidades de transição p_{ij} são aplicadas sempre que voltar ao estado inicial, não importa o que aconteceu no passado. Portanto a probabilidade de ocorrer X_{n+1} depende somente de X_n. De acordo com Bertsekas e Tsitsiklis (2002) um modelo de cadeias de Markov é especificado por meio da identificação de:

- um conjunto de estados $S = 1, ..., M$;
- um conjunto de transições possíveis, ou seja, os pares (i, j), para que $p_{ij} > 0$;
- os valores numéricos desses p_{ij} serem positivos.

As cadeias de Markov especificadas são uma sequência de variáveis aleatórias $X_0, X_1, X_2, ...X_n$, que tomam valores em S e que satisfaz a Equação 21.

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = p_{ij} \tag{21}$$

para todo n, todos os estados i, j pertencentes a S e todas as possíveis sequências i_0, ... , i_{n-1} de estados de anteriores.

Todos os elementos de um modelo de cadeias de Markov podem ser escritos como uma matriz de probabilidade, que é simplesmente uma matriz bidimensional cujo elemento na i-ésima linha e j-ésima coluna é p_{ij}. A matriz de probabilidade também pode ser lida como matriz de transição, cujos elementos são as probabilidades de transição de um estado para outro, como ilustrado na Figura 13.

$$\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}$$

Figura 13 – Matriz de Probabilidades

Com a finalidade de agilizar o entendimento dos processos, é de grande valia utilizar grafos (diagrama de transição) para expor o modelo da probabilidade de transição. O diagrama de transição é uma representação gráfica das cadeias de Markov. Neste diagrama são visualizados os estados, representados por círculos, e as probabilidades de transição entre os estados. A Figura 14 exemplifica uma matriz e um diagrama de transição com 3 estados.

$$\mathbf{P} = \begin{array}{c} \\ H \\ T \\ Q \end{array} \begin{array}{c} \begin{array}{ccc} H & T & Q \end{array} \\ \begin{pmatrix} P_{00} & P_{01} & P_{02} \\ P_{10} & P_{11} & P_{12} \\ P_{20} & P_{21} & P_{22} \end{pmatrix} \end{array}$$

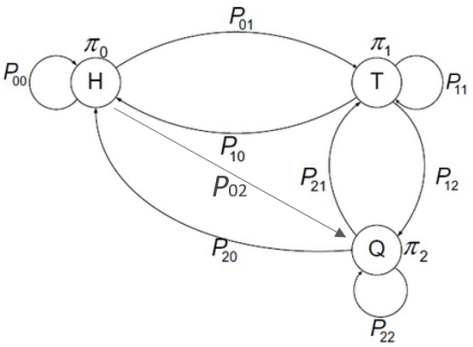


Figura 14 – Diagrama e Matriz de Transição

Muitos problemas que envolvem as cadeias de Markov exigem o cálculo de probabilidade em algum momento futuro condicionado ao estado atual. Esta probabilidade de transição de n-passos é definida pela Equação 22.

$$r_{ij}(n) = P(X_n = j | X_0 = i) \tag{22}$$

Ou seja, $r_{ij}(n)$ é a probabilidade de que após n períodos de tempo o estado seja j, dado que o estado atual é i. Isto pode ser calculado usando a recursão, conhecida como a equação Chapman-Kolmogorov. As probabilidades de transição de n-passos podem ser obtidos pela fórmula recursiva mostrada na Equação 23.

$$r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1) p_{kj} \tag{23}$$

Para $n > 1$, e todo i, j começado com: $r_{ij}(1) = p_{ij}$.

As cadeias de Markov podem ter diferentes variações quanto a tempo e a espaços de estados. Quanto ao tempo podem ser definidas como de tempo contínuo, possuindo um índice contínuo temporal, ou como de tempo homogêneo (ou cadeias de Markov estacionárias), são processos em que seguem a Equação 24.

$$Pr(X_{n+1} = x|X_n = y) = Pr(X_n = x|X_{n-1} = y) \tag{24}$$

Para todo n. A probabilidade da transição de n é independente.

Outra variação de cadeia de Markov são as cadeias de ordem m (ou uma cadeia de Markov com memória), onde m é finito. O estado futuro depende dos passados m estados. É possível construir uma cadeia (Y_n) de (X_n), que tem a propriedade de Markov "clássico", tendo como espaço de estado do m-tuplas ordenadas de valores X, ou seja, $Y_n = (X_n, X_{n-1}, ..., X_{n-m+1})$.

As cadeias de Markov possuem algumas propriedades, como define Meyn e Tweedie (2012). A Redutibilidade define que um estado j da cadeia é dito ser acessível a partir de um estado i se um sistema começou no estado i tem uma probabilidade diferente de zero de transição para o estado j em algum ponto. Logo, uma cadeia é irredutível se o seu espaço de estado é uma classe única comunicação, se é possível chegar a qualquer estado de qualquer estado.

A propriedade da Periodicidade afirma que um estado i tem período k se houver retorno ao estado i e deve ocorrer em múltiplos de passos de tempo k. Se $k = 1$, então o estado é dito ser aperiódico, pois o retorno ao estado i pode ocorrer em períodos irregulares. Caso contrário ($k > 1$), o estado é dito ser periódico com período k. A cadeia de Markov é aperiódica se cada estado é aperiódico. Uma cadeia irredutível só precisa de um estado aperiódico para implicar que todos os estados são aperiódicos.

A propriedade da Transitoriedade define que um estado i é dito transitório, se, uma vez que começa-se no estado i, existe uma probabilidade não nula de que nunca voltará a i. O estado i é recorrente (ou persistente) se não é transitório. Recorrência e transitoriedade são propriedades de classe, isto é, elas são válidas ou não de forma igual para todos os membros de uma classe comunicante.

A propriedade da Ergodicidade diz que um estado i é dito ser ergódico se ele tem uma recorrência aperiódica e positiva. Em outras palavras, um estado i é ergódico se for recorrente, tem um período de 1 e tem tempo de recorrência média finita. Se todos os estados em uma cadeia de Markov irredutível são ergódicos, então a cadeia é ergódica.

Quanto a análise de estado estacionário, como foi apresentado anteriormente, se a cadeia de Markov é uma cadeia de tempo homogênea, de modo que o processo é descrito por uma única matriz que independe do tempo p_{ij}, então o vetor π é chamado de distribuição estacionária (ou medida invariante) se todo $j \epsilon S$. Porém, a cadeia de Markov não precisa

ser necessariamente de tempo homogêneo para ter uma distribuição de equilíbrio, se há uma distribuição de probabilidade sobre estados π tal que satisfaçam a Equação 25.

$$\Pi_j = \sum_{i \varepsilon S} \Pi_i Pr(X_{n+1} = j | X_n = i) \tag{25}$$

Para cada estado j e cada tempo n, então π é uma distribuição em equilíbrio da cadeia de Markov. As cadeias homogêneas se caracterizam pelo fato de que a regra probabilística para obter X_{n+1} de X_n não depende do tempo n. Em certas situações esta hipótese é relaxada, permitindo então que as probabilidades de transição mudem com o tempo. Tais cadeias, cujas regras de transição dependem do tempo, são chamadas de cadeias não-homogêneas. Tal situação pode ocorrer em métodos de MCMC, ou seja, em situações em que um número de diferentes matrizes de transição são usadas, porque cada uma é eficaz para um tipo particular de mistura, mas cada matriz respeita uma distribuição de equilíbrio partilhada (BRÉMAUD, 2013).

3.5.1 Cadeia de Markov de Monte Carlo

O Método de Monte Carlo (MMC) é uma técnica estatística baseada em amostragens aleatórias massivas para obter resultados numéricos, isto é, repetindo sucessivas simulações um elevado número de vezes, para calcular probabilidades heuristicamente, tal como se, de fato, se registrassem os resultados reais em jogos de cassino, por exemplo. Este tipo de método é utilizado em simulações estocásticas com diversas aplicações. O MMC tem sido utilizado há bastante tempo como forma de obter aproximações numéricas de funções complexas em que não é viável, ou é mesmo impossível, obter uma solução analítica ou, pelo menos, determinística (HROMKOVIČ, 2013).

Em suma, pode-se dizer que o MMC realiza uma série de simulações com probabilidades aleatórias e ao fim define uma média dessas simulações visando a um objetivo que reflita a realidade desejada. Os MMC de passeio aleatório compõem uma grande subclasse das MCMC.

A MCMC surge como um método alternativo para solução de problemas complexos em inferência estatística (clássica e bayesiana) (ROBERT; CASELLA, 2013). A proposta é simular π via construção de uma cadeia de Markov em um conjunto objetivo tendo π como única distribuição estacionária. Os métodos MCMC garantem que, após um tempo suficientemente longo de simulação, elementos do conjunto objetivo podem ser amostrados com distribuição aproximadamente igual a π (NASCIMENTO, 2009).

A abordagem habitual da teoria de cadeias de Markov inicia-se com a função de transição da cadeia, definida por $P(x, y) \forall x, y \epsilon S$. A função de transição denota a probabilidade de a cadeia mover-se para o estado y dado que se encontra no estado x no tempo anterior.

O maior interesse da teoria de simulação de MCMC é determinar sob quais condições existe a distribuição estacionária π e quais condições fazem com que a função de transição

convirja para a distribuição estacionária. Sabe-se, da discussão apresentada na seção anterior, que a distribuição estacionária satisfaz $\pi = \pi P$ e que $P^n(x, y)$ representa a função de transição em n-passos da cadeia. Deste modo, deseja-se que $\lim_{n\to\infty} P^n(x, y) = \pi(y)$, isto é, quando n tende ao infinito, elementos da função de transição possam ser amostrados de uma distribuição aproximadamente igual à distribuição estacionária (NASCIMENTO, 2009).

Para gerar amostras de $\pi(x)$, os métodos MCMC encontram e utilizam a função de transição $P(x, y)$ que converge para $\pi(x)$ na n-ésima iteração. O processo é iniciado em um estado arbitrário x e, após um número suficientemente longo de simulação, as observações geradas são aproximadamente iguais a distribuição alvo $\pi(x)$. O problema então se resume em encontrar uma função de transição $P(x, y)$ apropriada.

As MCMC são métodos de simulação baseados em cadeias de Markov ergódicas (sistemas dinâmicos) onde a distribuição estacionária do processo estocástico é a distribuição *a posteriori* de interesse. Existem diversos tipos de métodos/algoritmos MCMC (BRÉMAUD, 2013), dentre estes:

- ❏ Algoritmo Metropolis–Hastings: esse método gera um passeio aleatório usando uma densidade proposta e um meio para rejeitar alguns dos movimentos propostos;
- ❏ Amostragem de Gibbs: este método requer que todas as distribuições condicionais da distribuição alvo sejam amostradas com exatidão. A amostragem de Gibbs é popular porque não requer uma distribuição proposta;
- ❏ Amostragem Slice: este método depende do princípio de que se pode amostrar a partir de uma distribuição por amostragem uniforme da região sob o gráfico da sua função de densidade. Alterna amostragem uniforme na direção vertical com amostragem uniforme da "fatia" horizontal, definida pela posição vertical atual;
- ❏ Múltipla tentativa Metropolis: este método é uma variação do algoritmo Metropolis-Hastings que permite múltiplas tentativas em cada ponto. Ao tornar possível dar passos maiores em cada iteração, ajuda a resolver o problema da dimensionalidade.
- ❏ Salto reversível: este método é uma variante do algoritmo Metropolis-Hasting, que permite propostas que alteram a dimensionalidade do espaço. O algoritmo de salto reversível é útil ao fazer a amostragem de MCMC, ou Gibbs, sobre modelos Bayesianos não paramétricos.

O Algoritmo de Metropolis-Hastings e Algoritmo Amostragem de Gibbs são os mais utilizados. O Algoritmo de Metropolis-Hastings é muito utilizado por ser pouco restritivo com relação a distribuição *a posteriori*. Pois para o uso deste algoritmo, é suficiente e necessário ter apenas a distribuição *a posteriori* a menos de uma constante de proporcionalidade e escolher uma distribuição proposta adequada, o que garante uma taxa de

rejeição, pois alguns valores das distribuição proposta serão rejeitados. O algoritmo de Gibbs é mais restritivo, pois para seu uso, é necessário conhecer as distribuições condicionais completas. Entretanto, não é preciso escolher uma distribuição proposta e com isso não existe taxa de rejeição.

Métodos mais sofisticados usam várias maneiras de reduzir a correlação entre amostras sucessivas. Esses algoritmos podem ser mais difíceis de implementar, mas eles geralmente exibem uma convergência mais rápida (ou seja, menos etapas para um resultado preciso). O algoritmo Monte Carlo Híbrido (MCH) tenta evitar o comportamento de caminhada aleatória introduzindo um vetor de momento auxiliar e implementando a dinâmica hamiltoniana[2], de modo que a função de energia potencial é a densidade alvo. As amostras de momento são descartadas após a amostragem. O resultado final do MCH é que as propostas se movem pelo espaço amostral em passos maiores. Eles são, portanto, menos correlacionados e convergem para a distribuição-alvo mais rapidamente (NEAL et al., 2011).

O Teorema de Perron-Frobenius define que em uma simulação de caminhada aleatória com matriz de transição P, a probabilidade de longo prazo de que o indivíduo que faz a caminhada esteja em um certo estado é independente do estado em que essa caminhada começou, e é definida pela distribuição estacionária. Ou seja, a caminhada aleatória "ignora" o passado (CHANG; PEARSON; ZHANG, 2008).

O conjunto de aplicações dos modelos que envolvem as cadeias de Markov é amplo. Eles incluem praticamente qualquer sistema dinâmico cuja evolução ao longo do tempo envolve incerteza, desde que o estado do sistema seja adequadamente definido (JUNIOR et al., 2015). Na estatística bayesiana, o recente desenvolvimento dos métodos de MCMC tem sido um passo fundamental para tornar possível o cálculo de grandes modelos hierárquicos que exigem integrações sobre centenas ou mesmo milhares de parâmetros desconhecidos (BRÉMAUD, 2013). Além da importância já descrita que as cadeias de Markov exercem para a modelagem e simulação, os processos de Markov são fundamentais para a implementação de um outro tipo de modelo, os ACE, um método poderoso para modelagem geográfica e temporal de sistemas complexos.

3.6 Autômatos Celulares

Como visto na Seção 3.4, sistemas complexos são uma classe de sistemas que são compostos de várias partes que interagem com a habilidade de gerar novas propriedades de forma dinâmica, geralmente envolvendo não linearidade. Ou seja, são sistemas que são sensíveis a mudanças em suas diferentes variáveis envolvidas. Estes tipos de sistemas são

[2] É um método de Monte Carlo para cadeias de Markov que visa a obtenção de uma sequência de amostras aleatórias a partir de uma distribuição de probabilidades para a qual a amostragem direta é difícil.

difíceis de serem analisados e estudados por métodos clássicos da matemática e da física (MELOTTI, 2009).

Modelo matemáticos de equações diferenciais são muito utilizados numa tentativa de "discretizar" sistemas complexos. As equações diferenciais são usadas para construir modelos matemáticos de fenômenos físicos. É uma equação cuja incógnita é uma função que aparece na equação sob a forma das respectivas derivadas. Dada uma variável x, função de uma variável y, a equação diferencial envolve x, y, derivadas de y e eventualmente também derivadas de x (AYRES, 1978), como ilustrado na Equação 26.

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + (x^2)y = 0 \tag{26}$$

Equações diferenciais têm propriedades intrinsecamente interessantes, nas quais uma solução pode existir ou não e caso exista, a solução pode ser única ou não, dado a incerteza por trás dos modelos mapeados. Quando determinados sistemas complexos são muito difíceis de serem modelados pelas equações diferenciais, pode-se aplicar abordagens com AC (CASTRO; LIMA, 2013). Ermentrout e Edelstein-Keshet (1993) mostram como AC podem representar equações diferenciais. Estas equações diferenciais surgem da tentativa de modelar-se complexos fenômenos físicos e biológicos através de simulações em computador.

Uma das abordagens para se estudar o comportamento de sistemas complexos é pelo uso de AC. Existem diversas aplicações no uso de AC, dentre elas, pode-se citar a modelagem de fenômenos naturais, físicos e biológicos. Além disso, os AC possuem alto nível de paralelismo quando implementados com programação paralela e executados em *hardwares* paralelos, fornecem alto desempenho computacional (CASTRO; LIMA, 2013).

De acordo com Wolfram (1983), um AC consiste de uma grade regular e uniforme com uma variável discreta em cada local (célula). A grade é vetorial se for unidimensional, ou denominada por "grid" se for bidimensional. As variáveis nas células podem adquirir qualquer valor dentro de um dado conjunto de valores possíveis. O estado do AC é definido pelos valores das variáveis em cada célula. Um AC evolui em unidades de tempo discretas, com o valor da variável em uma célula sendo afetado pelos valores das variáveis das células em sua vizinhança no momento anterior.

A vizinhança do AC é formada pelas células adjacentes e pela própria célula. As variáveis em cada célula são atualizadas simultaneamente com base nos valores de sua vizinhança no momento anterior e de acordo com um conjunto de regras locais. Estas regras podem ser determinísticas ou podem envolver elementos probabilísticos ou ruídos. Como exemplo, tem-se o procedimento mais simples em que o valor de uma célula, dado por uma regra determinística, pode ser revertido de acordo com uma certa probabilidade, com cada célula sendo tratada independentemente (LANZER, 2004).

Uma definição mais recente de Wolfram diz que:

> Um AC é uma coleção de células "coloridas" em uma grade de forma especificada que evolui por meio de várias etapas de tempo discretas, de acordo com um conjunto de regras baseadas nos estados das células vizinhas. As regras são então aplicadas iterativamente por quantas etapas de tempo, conforme desejado. (WOLFRAM, 2002)

Segundo Weimar (1997) um AC é caracterizado pelas seguintes propriedades fundamentais:

- ❏ Consistem em uma matriz, ou grade, de células;
- ❏ A evolução se dá em passos discretos de tempo;
- ❏ Cada célula é caracterizada por um estado pertencente a um conjunto finito de estados;
- ❏ Cada célula evolui de acordo com as mesmas regras que dependem somente do estado em que a célula se encontra e de um número finito de vizinhos;
- ❏ A relação com a vizinhança é local e uniforme.

O trabalho de Leite (2016) faz uma definição detalhada de AC, afirmando que podem ser definidos por uma quádrupla de elementos (G, V, Q, f) onde G é o espaço celular (normalmente uma rede regular, tal como Z ou Z^2), $V = i1, ..., is$ é a vizinhança (que é geralmente a mesma para todos os locais no espaço celular), Q é um conjunto finito de estados e f é uma função local de transição que associa um novo estado para cada configuração dos estados que compõem a vizinhança, como visto na Equação 27.

$$f : Q^s \rightarrow Q; (x1, ..., xs)\epsilon Q^s \rightarrow f(x1, ..., xs)\epsilon Q \tag{27}$$

em que $s = |V|$ é a cardinalidade do conjunto da vizinhança V (MANNEVILLE et al., 2012).

AC são definidos como a evolução dos estados das células que o compõem. O estado de uma célula $\theta_i^t \epsilon 0, 1$ indica que na posição i no tempo t a célula assume um dos estados definidos, neste caso 0 ou 1. Assumindo uma rede N-dimensional de células, tem-se um autômato N-dimensional. A função local de transição f_x , responsável pela evolução dos estados das células é definida como visto na Equação 28.

$$\theta_i^{t+1} = f(\theta_{i-k}^t, ..., \theta_i^t, ..., \theta_{i+k}^t) \tag{28}$$

Onde k é o índice de iterações. A função local de transição é aplicada simultaneamente em todas as células. O estado de uma célula no tempo $t + 1$ das $2k + 1$ células no tempo t, o que constitui sua vizinhança.

A vizinhança define o estado das células no próximo instante de tempo. As regras locais de transição ficam responsáveis por atualizar o valor de cada célula da rede, com

base nos valores das células que compõem a vizinhança do local, de acordo com o tipo de vizinhança e as condições de fronteira (LEITE, 2016), como ilustrado na Figura 15.

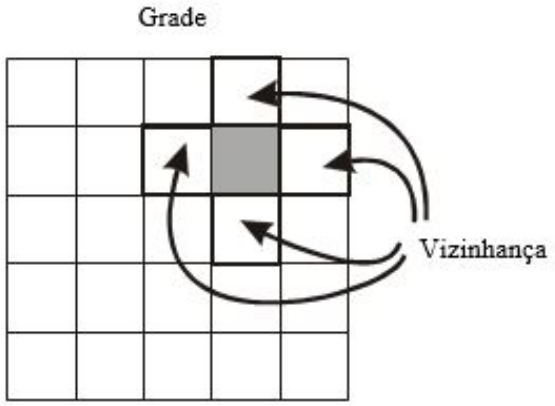


Figura 15 – Representação de uma grade, célula e sua vizinhança

Fonte: Castro e Castro (2015)

Um AC pode ser representado em diferentes dimensões, desde um vetor unidimensional até uma matriz N-dimensional, na qual cada célula é representada por uma posição. Um AC também pode possuir diferentes formas geométricas de células, como quadrados, triângulos e hexágonos. Na Figura 16 são ilustrados alguns exemplos de formatos de um AC e as células que o compõem (MANNEVILLE et al., 2012).

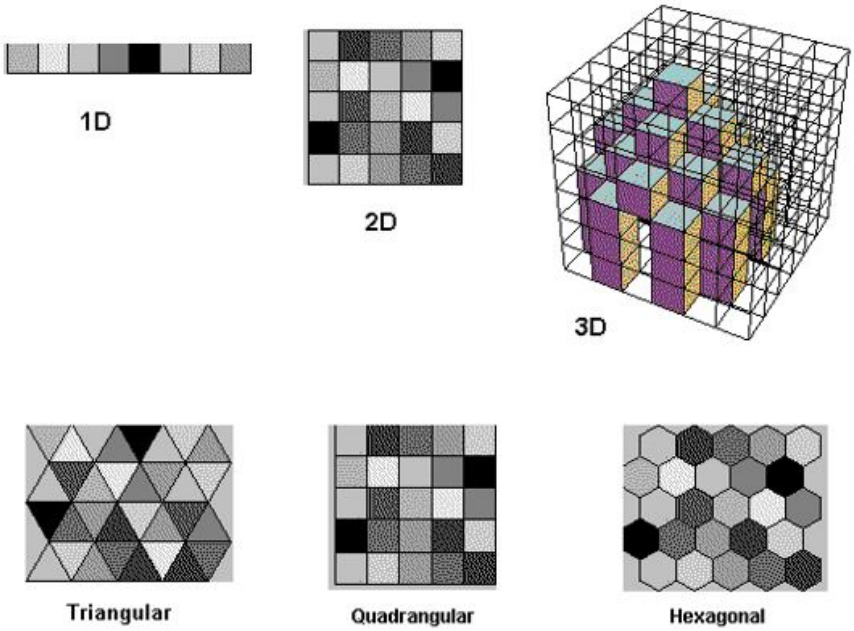


Figura 16 – Representação de diferentes dimensões de um autômato celular e de diferentes formatos de células

Fonte: Leite e Cerqueira (2014)

No caso de AC bidimensionais, é possível definir diferentes tipos de vizinhança tais como a de Von Neumann e a vizinhança de Moore. Na vizinhança de Von Neumann de raio unitário, cada célula é conectada às quatro adjacentes na vertical e na horizontal. A vizinhança de Von Neumann de raio r de uma célula, abrange as células dispostas ortogonalmente até a distância r desta célula. A vizinhançaa de Moore de raio r de uma célula é composta pela matriz quadrada de lado $2r + 1$ centrada na célula em questão. A vizinhança de Moore é composta das oito células adjacentes, considerando as células verticais, horizontais e diagonais (BATTY; COUCLELIS; EICHEN, 1997), como ilustrado na Figura 17.

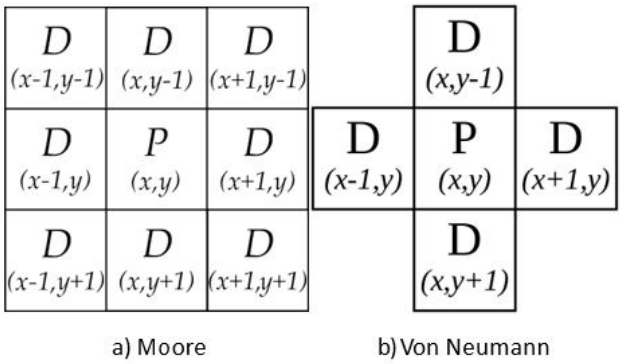


Figura 17 – Representação das vizinhanças de Moore e de Von Neumann

Fonte: Adaptado de Toffoli e Margolus (1987)

As bordas da grade do autômato podem ser preenchidas com zeros ou elas podem repetir os valores das células em posições opostas. Se as bordas são preenchidas com zeros, a grade é denominada uma ilha; se preenchida com as células das posições opostas, denomina-se torus (ou uma grade toroidal) (VARSHAVSKY; BAKAYEV, 1975), ilustrado na Figura 18.

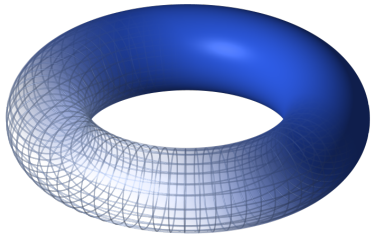


Figura 18 – Representação de um torus, ou grade toroidal

Pode-se dizer que os AC são compostos por cinco elementos fundamentais (LEITE, 2016)(LIU, 2008)(SEMBOLONI, 1997):

- Célula: é considerada a unidade básica do sistema. As células podem ser organizadas em um mosaico espacial que pode ser de uma, duas ou mais dimensões.

- Estado: define os atributos do sistema. Cada célula pode apresentar apenas um estado dentro de um conjunto de estados num determinado momento.

- Vizinhança: é o conjunto de células em que a célula em questão interage. Num espaço bidimensional existem duas tipologias de vizinhança: a vizinhança de von Neumann e a vizinhança de Moore.

- Regras de Transição: são um conjunto de condições ou funções que definem as alterações de estado de cada célula em resposta a seu estado atual e de seus vizinhos. O estado futuro de células é determinado pelas regras de transição em um período de tempo discreto.

- Tempo: especifica a dimensão temporal do AC em que os estados de todas as células são atualizados simultaneamente de modo iterativo ao longo do tempo.

Baseado nas características dos mapas espaço-temporais, Wolfram (1982) propôs uma classificação universal das regras, a saber:

- Classe 1: o estado global do autômato ao final da evolução é determinado com probabilidade 1 e independe do estado inicial. O comportamento do autômato é homogêneo, evoluindo para um ponto fixo (ciclo atrator contém um único estado) e em geral, gera no máximo duas bacias de atratores.

- Classe 2: o comportamento do autômato é denominado estável simples ou periódico. Os padrões gerados são constituídos por estruturas periódicas persistentes, com ciclos e transientes tipicamente curtos.

- Classe 3: o comportamento é caótico, não possuindo padrão reconhecível. A evolução é irregular. As regras pertencentes a essa classe possuem forte dependência das condições iniciais, apresentando grande instabilidade com relação a pequenas variações nos estados iniciais.

- Classe 4: o comportamento é complexo, gerando estruturas que evoluem imprevisivelmente. Logo, não é possível fazer predições acerca da evolução, sendo o comportamento determinado apenas por meio de simulações. Wolfram considera a possibilidade de que as regras pertencentes a essa classe possam apresentar a propriedade de computação universal - configurações iniciais adequadas podem especificar procedimentos algorítmicos arbitrários, fazendo com que o sistema funcione como um computador para aplicações gerais, capaz de avaliar qualquer função computável.

Wolfram (1984) foi o primeiro a demonstrar que um AC poderia exibir comportamento complexo mesmo com regras locais simples. Sua classificação demonstra que tais regras

podem levar a uma espécie de auto-organização, o que contribui inicialmente para uma maior compreensão do fenômeno de formação espontânea de padrões.

Uma característica dos modelos de AC é que comportamentos complexos que ocorrem ao longo de um espaço celular podem emergir da aplicação de regras simples e, assim, esses modelos permitem identificar, simular e observar o comportamento global de sistemas dinâmicos, ou seja, sistemas não completamente entendidos, mas para os quais seus processos locais são bem conhecidos (PARK; WAGNER, 1997). Tobler (1979) definiu representações de modelos de AC para representar geografias, como ilustrado na Figura 19. Pode-se observar que entre diferentes modelos que consideram o histórico para a determinação de um estado futuro, o modelo geográfico utiliza a vizinhança da célula de interesse para a determinação do estado futuro desta célula.

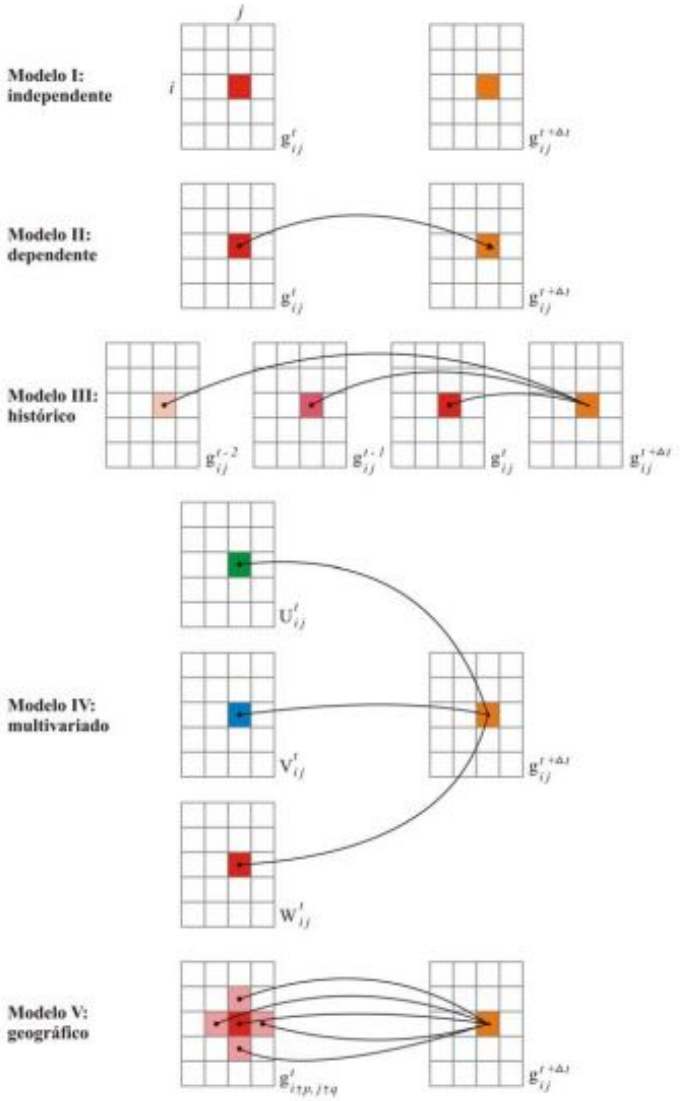


Figura 19 – Cinco tipos de modelos celulares para representações geográficas

Fonte: Tobler (1979)

3.6.1 Autômatos Celulares Estocásticos

Os AC são um sistema dinâmico de tempo discreto de entidades interagentes, cujo estado é discreto. O estado da coleção de entidades é atualizado a cada tempo discreto de acordo com alguma regra simples e homogênea. Os estados de todas as entidades são atualizados em paralelo ou em sincronia. ACE, ou probabilísticos, possuem regra de atualização estocástica, o que significa que os estados das novas entidades são escolhidos de acordo com algumas distribuições de probabilidade. É um sistema dinâmico aleatório de tempo discreto (FERNÁNDEZ; LOUIS; NARDI, 2018).

ACE são uma especialização de AC com uma aleatoriedade embutida em algum nível da evolução, representando uma fonte de incerteza do processo evolutivo. Pode ser que a célula a ser atualizada seja escolhida ao acaso; pode ser que a própria regra de transição local envolva probabilidades na atualização de cada célula ou, ainda, que a regra de transição seja escolhida ao acaso (LOZANO, 2017).

Os ACE são sistemas descritos por um conjunto de variáveis discretas onde os estados de cada célula são atualizados de forma síncrona e que obedecem a regras probabilísticas, dependendo apenas do estado de seus vizinhos no instante corrente (FERREIRA, 2009). Pelo fato de o estado corrente do autômato ser o único responsável pelo próximo estado global, a evolução desses sistemas pode ser modelada por uma cadeia de Markov homogênea finita, e o resultado final é a distribuição de probabilidades correspondente ao estado estacionário, obtida a longo prazo (LOZANO, 2017).

Como um processo de Markov em tempo discreto, um ACE é definido em um espaço de produto $E = \prod_{k \epsilon G} S_k$ (produto cartesiano), onde G é um gráfico finito ou infinito e S_k é um espaço finito, como por exemplo $S_k = -1, +1$ ou $S_k = 0, 1$ (FERNÁNDEZ; LOUIS; NARDI, 2018).

A regra de evolução se torna uma função booleana probabilística. Isto pode ser escrito como uma função booleana determinística de alguns bits aleatórios, permitindo assim o uso de codificação diversificada. Para isso, é necessário escrever as funções booleanas determinísticas que caracterizam as configurações na vizinhança com a mesma probabilidade (BAGNOLI, 1998).

Os ACE se apresentam como métodos relevantes para a modelagem de sistemas complexos onde há incerteza sobre as variáveis envolvidas nas unidades que constituem o sistema. O conhecimento não necessariamente exato, mas de um espaço de possiblidades de uma variável em um determinado local, ou período. é o suficiente para realizar simulação e buscar regras que se aproximem de uma modelagem computacional precisa da realidade do sistema complexo.

Capítulo 4

Metodologias e Propostas

Como já apresentado nas seções anteriores, a agricultura e as agroindústrias têm um importante papel no cenário econômico e social mundial. Porém, esses setores precisam se atentar a problemas quanto ao aumento da demanda mundial por alimentos, otimizando seus processos de produção, enquanto se preocupam com a preservação do meio ambiente. Visando a realização destas melhorias, o setor agropecuário começa a observar que o investimento em tecnologias e inovações acarreta ganhos para os seus processos produtivos.

Um dos principais fatores que possibilitou que o agronegócio se tornasse um setor estratégico para a economia brasileira foi o investimento na modernização da agropecuária, gerando um aumento da produtividade e da diversidade da cadeia agrícola (NETO; COSTA et al., 2005). O investimento em novas tecnologias possibilitou um conhecimento especializado relativo à produção, facilitando a tomada de decisão por partes dos gestores e a otimização dos processos produtivos.

A utilização de novas tecnologias na agricultura resultou na geração de uma grande massa de dados. A análise dos resultados e métricas colhidas da produção gerou ganhos para a economia desse setor. Porém, em nível de refino dessas massas de dados e de conhecimento específico e detalhado dos processos produtivos do agronegócio, ainda há muito a se investir (FILHO; FISHLOW, 2017).

O Brasil é um dos maiores representantes mundiais na produção e exportação de produtos agrícolas (FAO, 2018). Porém, tem em grande parte de sua agroindústria setores de Pesquisa e Inovação (P&D) emergentes ou não existentes. Resultados de pesquisas das últimas décadas demonstram que o investimento em novas tecnologias e inovações possibilitam ganhos sustentáveis para as indústrias, como as do setor agrário (FILHO; FISHLOW, 2017).

Logo, o que pode-se perceber durante o desenvolvimento desta pesquisa, tanto na literatura como em campo, foi que a agroindústria precisa modernizar os processos produtivos e um caminho que vem sendo muito utilizado é a utilização de sensores e atuadores nas plantações. No entanto, mesmo quando esses recursos são utilizados, há uma grande incerteza sobre o que fazer com os dados coletados.

Visando a problemática observada, este projeto busca contribuir oferecendo para a comunidade propostas de modelagem e simulação de monoculturas alinhadas com as tecnologias atuais de dispositivos móveis e IoT. Os modelos propostos surgem de um estudo que interliga as necessidades observadas em campo e com os modelos matemáticos e computacionais apropriados para tratarem os problemas de forma eficiente.

O estudo teve início com a realização de um mapeamento sistemático da literatura para um entendimento macro dos problemas enfrentados pelas agroindústrias e que métodos estão sendo utilizados para auxiliar o tratamento dessas problemáticas. O mapeamento sistemático seguiu os critérios metodológicos de Kitchenham (2004), como descrito no Capítulo 2. O mapeamento sistemático é muito importante para a definição de um estado da arte conciso das pesquisas e problemáticas atuais da área, porém, também se buscou uma observação da situação atual em campo, por meio de visitas as usinas, para uma melhor fundamentação da pesquisa.

A chamada quarta revolução industrial, ou Indústria 4.0, é impulsionada por avanços tecnológicos e de automação nas indústrias. O Brasil possui polos de inovação espalhados pelos seus estados e tem o apoio de empresas, como o Serviço Nacional de Aprendizagem Industrial (SENAI), que investem no mercado industrial para tornar a Indústria 4.0 uma realidade comum no país. Este projeto contou com uma imersão na realidade de um Instituto SENAI de Inovação (ISI), podendo observar de perto as tecnologias que estão sendo utilizadas e desenvolvidas em âmbito nacional. Pode-se perceber a importância que o uso da IoT, relacionada a modelos computacionais para a tomada de decisão, vem tendo para as indústrias.

Dado a temática atual e características específicas da problemática, a pesquisa envolveu imersões e entrevistas em diferentes áreas, empresas e setores. Além do ISI, houve a parceria de professores e pesquisadores do Departamento de Agronomia e de Engenharia Florestal da Universidade Federal Rural de Pernambuco (UFRPE).

Frente a todas as informações e *insights* coletados do mapeamento sistemático e da imersão em setores chaves pode-se ter uma definição da problemática a ser abordada, de possíveis soluções e de um processo metodológico a ser seguido. Com o que foi observado chegou-se à conclusão de que a modelagem computacional de monoculturas é um importante tema para as agroindústrias, visto que hoje muitas delas não possuem tecnologias de monitoramento de suas plantações. Então, visando a esta problemática, decidiu-se seguir o desenvolvimento de propostas de modelos para a realização de modelagem e simulação de monoculturas, em que de preferência, os agrônomos pudessem ter um acompanhamento em tempo real da situação dos cultivos.

Com a definição do desenvolvimento de modelos computacionais, surgiu a necessidade de se encontrar uma parceria para o levantamento dos dados e realização dos experimentos, testes e validações dos modelos. Em conversa com os especialistas foi relatado que só a existência de modelos empíricos e genéricos para a modelagem de monoculturas já seria

interessante, porém a validação destes é essencial para um trabalho científico.

Para a validação da plataforma proposta neste trabalho, optou-se pela modelagem de monoculturas de cana-de-açúcar, justificado pelo fato do Brasil ser o principal produtor mundial e pela relevância e leque de produtos derivados deste cultivo (FAO, 2018). Outro fator importante é o de que o IBGE afirma que lidar com culturas de cana-de-açúcar é algo complexo (IBGE, 2018). Ou seja, o objetivo deste trabalho é lidar com a modelagem de sistemas complexos e segundo especialistas da área, nada mais complexo no nosso ambiente do que monoculturas de cana-de-açúcar. Cultivos de tomates, por exemplo, possuem variáveis de interferência mais fáceis de serem monitoradas por modelos mais simples (OLIVEIRA; VIANNA, 2015).

Portanto, para se alcançar os objetivos deste trabalho foi importante de início a definição de modelos genéricos para as modelagens e simulações de monoculturas e a imersão para a adaptação dos modelos para serem utilizados em um cenário real. Para a imersão, este estudo utilizou o *Toolkit* HCD como processo metodológico para conseguir seus objetivos de entregar uma solução alinhada com as necessidades dos *stakeholders*. Para a adaptação dos modelos escolhidos é fundamental a análise dos dados de onde se deseja aplicar as soluções propostas. Para a MD das bases fornecidas pela usina do estudo de caso, utilizou-se as etapas do CRISP-DM. A Figura 20 ilustra o percurso metodológico percorrido no decorrer desta pesquisa.

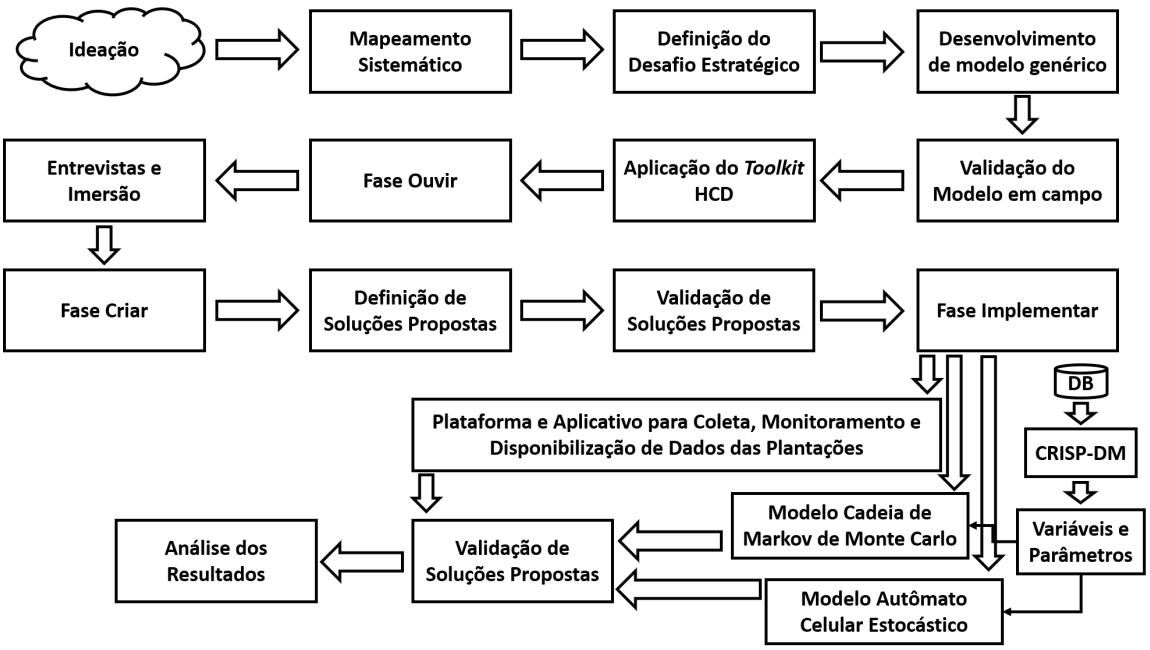


Figura 20 – Percurso metodológico da pesquisa

A seguir são descritos o emprego do *Toolkit* HCD no estudo de caso, a utilização do CRISP-DM no processo de MD e ainda são descritas as principais ferramentas utilizadas, arquiteturas e fluxogramas de pesquisa.

4.1 *Toolkit* HCD

O Design Centrado no Usuário, *human centered design*, ou HCD, ajuda as organizações a se relacionarem melhor com as pessoas às quais serve, transforma dados em ideias implementáveis, facilita a identificação de novas oportunidades e aumenta a velocidade e eficácia na criação de novas soluções (IDEO, 2014). O *Toolkit* HCD é uma metodologia para o desenvolvimento de soluções inovadoras focada nas necessidades dos usuários. O seu processo se inicia com a definição de um desafio estratégico específico e prossegue por três fases principais, são elas: Ouvir, Criar e Implementar, como ilustra a Figura 21.

As melhores soluções são aquelas construídas colaborativamente, através do somatório de diferentes perspectivas e visões(VIANNA, 2012). Estudos demonstram que o uso dessa metodologia tem alcançado bons resultados na criação inovadora de produtos, projetos, modelos e serviços, sejam eles educacionais, sociais ou mercadológicos (PREECE; ROGERS; SHARP, 2015)(BROWN, 2008)(VIANNA, 2012).

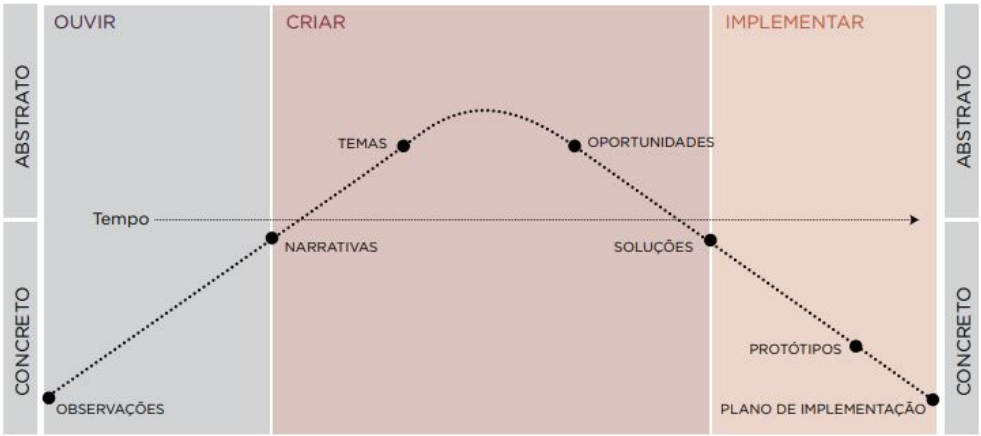


Figura 21 – Processo do HCD

Fonte: IDEO (2014)

Esse trabalho contempla todas as etapas do processo de desenvolvimento baseado no HCD, desde a imersão, passando pela ideação, prototipação e concluindo o processo com a implementação e validações de uma solução proposta. Visando orientar o processo do *Toolkit* HCD, precisa-se definir um Desafio Estratégico (DE). O DE proposto para esse trabalho consiste em responder a seguinte pergunta: "Como fazer com que os agrônomos e pesquisadores agrários consigam acompanhar e realizar simulações de seus canaviais em tempo próximo ao real?". A seguir será apresentado e discutido como o desenvolvimento do estudo de caso desta pesquisa aplicou cada uma das fases do *Toolkit* HCD.

4.1.1 Fase Ouvir

Esse trabalho inicia-se com a fase Ouvir, em que os principais objetivos são: determinar quem deve ser abordado, ganhar empatia com os usuários chaves e coletar histórias. É

importante começar a fase Ouvir pela etapa de "Avaliar o conhecimento preexistente", na qual destaca-se os saberes que já se possui com base no DE definido.

Este projeto é uma pesquisa do Departamento de Estatística e Informática (DEINFO) da UFRPE, tendo como área base a ciência da computação. Logo, já se tem conhecimentos técnicos em algoritmos e programação e dessa forma, a hipótese inicial foi desenvolver uma solução de *software* para a modelagem e simulação das monoculturas atendendo aos requisitos que fossem apontados pelos *stakeholders*. Porém, para que esse desenvolvimento fosse possível, teve que se extrair muitos conhecimentos técnicos das ciências agrárias, do mapeamento sistemático realizado e de entrevistas com especialistas.

Após elencar os conhecimentos preexistentes, foram levantadas perguntas chaves para serem feitas com o público-alvo dos modelos propostos. Dentre as perguntas elaboradas, tem-se: (I) O que fazem, pensam ou sentem os *stakeholders* em relação ao conhecimento e monitoramento dos canaviais? (II) O que os membros da usina acham das ofertas de soluções disponíveis no contexto de modelagem e simulação de canaviais? (III) Quais os principais desafios para implementar as ideias propostas nesta pesquisa? (V) Onde estão as maiores necessidades observadas? (VI) Como a estratégia de entrevistas deve ser planejada?

Na etapa de "Identificar pessoas com quem conversar", fez-se necessário identificar o público-alvo do projeto. Foram identificados três tipos distintos de perfis: membros ideais, membros não ideais e membros medianos. Os membros ideais são pessoas que demonstram comportamentos desejáveis e alinhados aos pensados para o projeto. Os membros não ideais são pessoas mais resistentes as soluções sugeridas. E os membros medianos são pessoas que estão no meio termo entre os dois grupos anteriores.

Em relação aos membros ideais foram considerados pesquisadores e agrônomos que já lidam com inovação e tecnologia no seu cotidiano. Já os membros medianos foram aqueles que não utilizam, ou lidam, com inovação e tecnologia com frequência, mas consideram esses recursos úteis para algumas áreas. Por fim, os membros não ideais selecionados foram aqueles que se auto consideravam conservadores e que acreditam que há formas melhores de se atingir os objetivos do DE do que com o uso da computação. As entrevistas foram realizadas com integrantes da Usina São José, do Departamento de Agronomia da UFRPE, do Departamento de Engenharia Florestal da UFRPE e *stakeholders* identificados na rede social LinkedIn.

Na etapa de "Escolher métodos de pesquisa", o presente trabalho utilizou: entrevistas individuais; entrevistas em grupo e a auto-documentação, sendo todos previstos pelo *Toolkit* HCD. Eles foram utilizados em uma abordagem casual com os entrevistados, em que os conceitos de "conversa" e "opinião" definiram as técnicas utilizadas.

As entrevistas individuais foram realizadas em reuniões presenciais e por formulário *online* encaminhado para pessoas específicas e disponibilizado em grupos que atendessem ao perfil do projeto. A Tabela 4 mostra os questionamentos realizados.

Tabela 4 – Questionamentos realizados na entrevista.

I	Porque você [não] acha interessante o monitoramento dos canaviais?
II	Porque você [não] acha interessante a realização de simulações de cenários dos canaviais?
III	Conte como você faz para monitorar e realizar simulações (predições) de cenários dos canaviais.
IV	Porque você [não] utiliza um sistema para obter essas informações?
V	Fale como é utilizar o seu recurso atual (se existir) para monitorar e simular cenários dos canaviais.
VI	O recurso utilizado atualmente permite um monitoramento georreferenciado em tempo real e a distância?
VII	Seria interessante a utilização de *smartphones* e *tablets* como ferramentas para auxiliar as coletas de informações das plantações?

Todas as entrevistas em grupo ocorreram na UFRPE. Já as entrevistas com os *experts* se deram em maior parte por reuniões presenciais, trocas de e-mail e ligações. Dentre os especialistas que colaboraram com o desenvolvimento deste projeto destaca-se o engenheiro agrônomo e coordenador da Usina São José Dr. Antônio Lima e o professor do Departamento de Agronomia da UFRPE Dr. Emídio Cantídio, especialista na área de fertilidade do solo.

De acordo com IDEO (2014), a etapa "Desenvolver a abordagem de entrevista" é formada por três abordagens, são elas: Guia de Entrevista; Conceitos Sacrificiais e Técnicas de Entrevista. Nessa etapa, procurou-se trazer o tom de casualidade às pesquisas, geralmente discorrendo sobre um tema padrão e seguindo uma linha de conversa pré-estabelecida para observar as respostas dos entrevistados. O contexto da conversa era iniciado com os seguintes tópicos: conhecimentos sobre o canavial e técnicas aplicadas; quais as maiores necessidades e problemas enfrentados. A expansão da conversa dava-se com perguntas como: "Seria interessante que existisse algo para resolver/ajudar nessa problemática?"; "Um *software* para *smartphones* e *tablets* poderia ser útil? E para computadores?". A sondagem de profundidade teve perguntas como: "Se esse *software* existisse como seria?"; "Como você o acharia interessante?".

Na etapa "Desenvolver um modelo mental", os preceitos de "Mente de principiante" e de "Observar x Interpretar" foram seguidos como regra para não influenciar os entrevistados durante toda a fase Ouvir. Sempre que havia uma fuga ao foco no processo de pesquisa, utilizava-se de técnicas de condução, sem influenciar o grupo, ou indivíduo.

4.1.2 Fase Criar

A segunda fase do *Toolkit* HCD é a Criar. Conforme IDEO (2014), dentre os objetivos dessa fase, tem-se: entender os dados, identificar padrões, definir oportunidades e criar soluções. O *Toolkit* HCD alerta que para transformar pesquisas em soluções para o mundo real é preciso filtrar e selecionar informações, traduzindo *insights* sobre a realidade em

oportunidades para o futuro. Assim, adota-se um ponto de vista generativo para criar diversas soluções em *brainstorms* e rapidamente converter algumas delas em protótipos.

Para iniciar a fase Criar é preciso "Desenvolver uma abordagem", o que consiste em desenvolver um entendimento profundo sobre a problemática e traduzi-lo em inovações. Decidiu-se pela utilização da abordagem participativa e a abordagem empática. Durante toda a fase Criar, as duas abordagens foram utilizadas procurando a imersão dos *stakeholders* no processo de criação de uma proposta de solução. Suas necessidades, gostos e preferências foram considerados para a definição do projeto proposto.

A etapa "Compartilhando Histórias" consiste em transformar as histórias coletadas durante a fase Ouvir, e transformá-las em dados e informações que serão utilizados para inspirar a criação de oportunidades, ideias e soluções. Diante das histórias ouvidas, se traçou uma estratégia onde essas histórias deveriam ser transformadas em informação, mas sem perder o contexto. Assim, as respostas obtidas nos questionários, bem como os depoimentos das entrevistas, foram analisados visando sintetizar as histórias semelhantes e tópicos importantes. A maioria das histórias levantadas nessa etapa foram dos públicos ideais e medianos, sendo as considerações dos não ideais levadas em conta para reforçar a solução e planejar possíveis melhorias.

Pode-se perceber que muitas histórias foram repetidas, no geral foi observado que os grandes problemas enfrentados são: (I) Como prever a safra?; (II) Como simular cenários?; (III) Como monitorar a plantação?; (IV) Como analisar os dados coletados? Os entrevistados normalmente utilizam planilhas em Excel para fazer as análises, mas essas normalmente são mais superficiais. Acham interessante que houvesse uma solução para *tablets* e *smartphones*, porém alertam que nas plantações normalmente não a sinal de internet ou de telefonia. A etapa de "Encontrar temas" consiste em explorar as relações entre informações. Dessa forma, as principais problemáticas destacadas pelos entrevistados foram selecionadas como temas chave. A etapa "Criando áreas de oportunidade" refere-se ao processo de traduzir *insights* em oportunidades. Para o desenvolvimento das áreas de oportunidade desse trabalho, mais uma vez foram usadas como base as histórias planejadas anteriormente. Frente a elas se elaborou as distintas áreas de oportunidade de acordo com diferentes *insights*.

A etapa de "Brainstorm de novas soluções" possibilita pensar de forma abrangente e sem qualquer restrição. Em colaboração com os *stakeholders*, criou-se um documento compartilhado com todas as principais informações já obtidas, que foram consideradas de maior relevância para o projeto. Frente às histórias, *insights*, temas e áreas de oportunidade selecionados, teve-se que ser criativo e pensar em ideias para atacar os problemas identificados das mais diversas formas possíveis.

Ao fim dessa tempestade de ideias destacou-se algumas propostas: (I) Um aplicativo para que os agrônomos possam inserir dados georreferenciados da plantação através de um *tablet* ou *smartphone* offline. Esses dados posteriormente são sincronizados em nu-

vem com outros dados da rede IoT; (II) Um modelo computacional para predição de safras baseado em cadeias de Markov; (III) Um modelo computacional de AC para a realização de simulações de cenários dos canaviais. (IV) Uma plataforma em nuvem para monitoramento em tempo real das plantações via dispositivos móveis.

Ao analisar a Figura 22 e comparar com a Figura 15, percebe-se que as plantações com seus talhões e distribuições seguem um desenho semelhante aos observados nos AC 2D, com sua grade, células e vizinhanças.

Ao analisar as propostas destacadas, bem como seus aspectos positivos e negativos, e as habilidades técnicas existentes, chegou-se a proposta de solução de uma plataforma que facilitasse o dia a dia dos pesquisadores e agrônomos da usina. Sendo possível interligar modelos de coleta, processamento e simulação de dados, visando ao monitoramento dos canaviais e otimização das safras.

Figura 22 – Plantações de cana-de-açúcar

A etapa "Transformando ideias em realidade" consiste em construir protótipos e validações, para que ideias se tornem tangíveis e sejam testadas e avaliadas por outros de forma rápida e barata, antes que se apegue a uma ideia específica. Por se tratar de um estudo científico, a ideia desse projeto é que para cada contribuição proposta se tenha a publicação de um artigo, para a validação individual das ideias.

Foram desenvolvidos protótipos, arquiteturas e modelos separadamente, visando aos problemas destacados e que futuramente essas soluções possam ser integradas em uma única plataforma. Houve a preocupação de que nos testes com os *stakeholders* se tivesse uma ideia de *storyboards* e usabilidade completa da plataforma, permitindo uma melhor coleta de *feedbacks*. Logo, como resultado desta etapa, foram propostos:

- ❏ Um aplicativo para dispositivos móveis com o intuito de coletar dados e realizar o monitoramento georreferenciado e em tempo próximo ao real da plantação (Figura 23);

- ❏ Uma arquitetura em nuvem para armazenamento e compartilhamento dos dados das plantações oriundos do aplicativo e dos sensores existentes nas plantações;

- ❏ Uma adaptação do modelo de cadeia de Markov para o cenário de predição de safras;
- ❏ Uma adaptação do modelo de AC para a simulação de cenários futuros das plantações.

Ao apresentar protótipos para o público-alvo, obtém-se *feedbacks* honestos com maior facilidade, visto que há espaço para comparações e contrastes de ideias. Com isso, a última etapa da Fase Criar é "Coletando *Feedback*", em que o foco é apresentar os protótipos ao público-alvo para obtenção de *feedbacks* qualitativos e quantitativos, a fim de escolher e ajustar as soluções propostas com as sugestões obtidas. Dois grupos foram abordados durante a coleta de *feedbacks*, os grupos das pessoas que já tinham sido abordados na Fase Ouvir, como também um novo grupo, devido a necessidade de encontrar opiniões diferentes. O novo grupo também foi formato por agrônomos e pesquisadores da usina e da UFRPE.

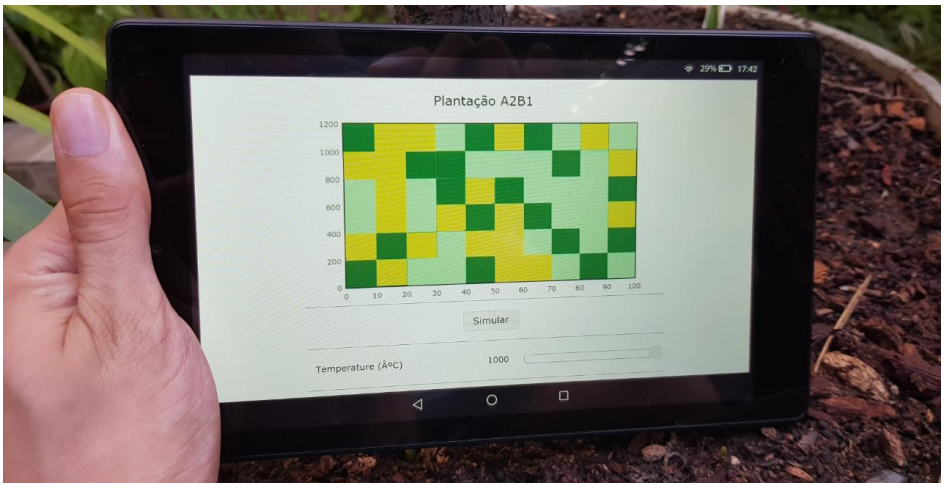


Figura 23 – Protótipo do aplicativo para monitoramento de setores do canavial

Nesta etapa, procurou-se respostas para questões, as quais são encontradas na Tabela 5. Nenhuma dessas perguntas foi realizada de maneira direta (como se seguisse um questionário), para não prejudicar a qualidade dos *feedbacks*, elas foram obtidas frente a observações dos usuários.

Tabela 5 – Questionamentos realizados na fase de *feedbacks*.

I	As soluções propostas proporcionam um monitoramento eficiente das plantações?
II	As soluções propostas possibilitam a predição de safras?
III	As soluções propostas possibilitam a simulação de cenários futuros das plantações?
IV	As propostas são atrativas?
V	Você utilizaria as soluções propostas no seu dia a dia?
VI	O público não ideal utilizaria a solução proposta?
VII	As soluções propostas facilitariam e otimizariam o seu cotidiano de trabalho?
VIII	O que está ruim?
IX	O que adicionaria?

Inicialmente buscou-se os grupos já entrevistados na fase Ouvir, porém nem todas as pessoas ouvidas durante essa fase responderam a solicitação para a validação das soluções propostas (detalhes serão citados no Capítulo 7). Portanto, se buscou outros grupos de usuários para testes, de preferência que fossem personas dos grupos ideais ou medianos, para enriquecerem as propostas. Os usuários não ideais também foram essenciais nessa etapa, pois quando as melhorias propostas por esse grupo não entravam em choque com os desejos dos demais usuários, essas eram acatadas.

Para a realização dos testes das soluções propostas foram disponibilizados instaláveis do aplicativo para serem utilizados na usina. A arquitetura em nuvem proposta foi documentada e enviada por e-mail. Os modelos de cadeia de Markov e de AC foram ilustrados em *storyboards*[1] e tiveram suas funcionalidades documentadas. Procurou-se deixar os usuários testarem e criticaram as propostas, sem interferir. Parte da obtenção dos *feedbacks* foi feita de forma remota, em que os usuários enviaram suas opiniões por e-mail ou mensagens em redes sociais, e outra parte foi obtida de forma presencial, em entrevistas individuais, ou em grupo.

Os *feedbacks* obtidos foram de extrema importância para a qualidade final do projeto, pois foram determinantes para ajustes de viabilidade de precisão dos modelos. Entre os principais ajustes realizados se teve: (I) O aplicativo deve funcionar offline, armazenando dados em *localstorage*, e quando se conectar à internet, deve sincronizar os dados armazenados localmente com a nuvem; (II) Só uma simulação dos modelos de cadeia de Markov e de AC não é o suficiente para um resultado preciso, daí surge a ideia de se utilizar MCMC e ACE; (III) As variáveis chaves e especificações da cadeia de Markov e principalmente do AC devem ser oriundas não somente da MD da usina, mas devem ser ajustadas por conhecimentos dos especialistas. Detalhes sobre a etapa de MD estão presentes na Seção 4.2.

4.1.3 Fase Implementar

Após o desenvolvimento e coleta de *feedbacks* das soluções propostas, inicia-se a fase de implementação das soluções, a fim de torná-las factíveis. Os objetivos da fase Implementar são: identificar capacidades necessárias; criar um modelo financeiro sustentável; desenvolver a sequência de projetos de inovação; criar pilotos e medir impactos (IDEO, 2014).

A fase implementar começa com o desenvolvimento de um modelo de receita sustentável. O sucesso das soluções a longo prazo depende do desenvolvimento intencional de uma estratégia de rentabilidade que possa sustentar a oferta. Dessa forma, tem-se como proposta de valor para os usuários a entrega de conhecimentos refinados sobre as plantações, podendo monitorá-las em tempo próximo ao real, realizar simulações e compartilhar

[1] Sequência de telas que representam o *layout* e a jornada do usuário na utilização das soluções propostas.

dados entre os pesquisadores com facilidade, tudo isso resultando em informações relevantes para a tomada de decisão e otimização das safras. A fonte de receita para garantir a sustentabilidade da plataforma deve ser oriunda de investimentos da usina e órgãos fomentadores de P&D.

Posteriormente, chegou-se na etapa de "Identificação de capacidades necessárias para implementar soluções". As capacidades requeridas para a realização do projeto vão desde os conhecimentos técnicos em programação e algoritmos até os conhecimentos específicos das ciências agrárias e peculiaridades do cultivo da cana-de-açúcar. O aplicativo foi desenvolvido utilizando tecnologias web (HTML5, Javascript e CSS3), o que possibilita que seu *deploy*[2] possa ser realizado para diferentes sistemas operacionais e dispositivos. Os serviços do aplicativo são ligados a *endpoints*[3] hospedados na AWS da Amazon, que provê um serviço desenvolvido em Node.js e com banco de dados MongoDB que armazenam e disponibilizam os dados coletados da plantação, sejam esses oriundos do aplicativo ou dos sensores existentes.

Para o desenvolvimento do modelo de cadeia de Markov de Monte Carlo utilizou-se a linguagem de programação Python. Já o modelo de ACE utilizou-se um *back-end* em Python e um *front-end* com tecnologias web (HTML5, Javascript e CSS3), garantindo uma visualização mais usual da grade e das células do autômato. Todos os algoritmos desenvolvidos foram idealizados tomando por base conhecimentos extraídos da literatura, da MD fornecidos pela usina e das entrevistas e ajustes com os especialistas. Através disso foi possível a identificação de variáveis chaves, ciclos temporais e classificações dos cultivos.

Também é fundamental que os dados e informações fornecidos pelos modelos tenham integridade e confiabilidade confirmadas. Para isso, os modelos utilizaram bases de dados fornecidas pela usina como base, mas também permitem que os pesquisadores ajustem parâmetros das modelagens antes de realizar as simulações. Os especialistas, caso achem necessário, podem ajustar as probabilidades de transição manualmente nas matrizes de transição e fatores de propagação, por exemplo, em ambos os modelos.

Quanto às necessidades financeiras para a implantação da solução, teve-se que investir em planos de serviços da AWS, em uma infraestrutura refinada de sensores nas plantações e em *tablets* e computadores para executarem os *softwares*. Ambos os modelos não necessitam de alto poder computacional para executarem as simulações, onde podem ser realizadas na nuvem e só as requisições e exibição dos resultados serem processadas nos dispositivos dos usuários.

Na etapa "Planejando um conjunto de soluções" estabeleceu-se dois marcos: a fase em que se encontra o projeto, e para onde o mesmo poderá avançar. Em suma, as soluções propostas encontram-se implementadas separadamente, não há uma integração entre os

[2] Disponibilização do *software* em uma versão executável.

[3] Conjunto de rotas de um *web service* que possibilita interações do cliente com uma API disponibilizada.

modelos. Os agrônomos precisam executar a aplicação em Python de cadeias de Markov para ter uma predição de safras e executar a aplicação de autômatos para simular cenários. Espera-se que futuramente as soluções sejam unificadas em uma única plataforma, onde em um só lugar os pesquisadores e agrônomos possam acompanhar as plantações, inserir novos dados e realizar simulações de safras e cenários.

Na "Criação de um calendário de implementação" foi desenvolvido um plano de ações. Este existe desde o começo do projeto, onde cada entrega de artefato é marcada como um *strike*, caso tenha sido realizada no período pré-definido. Na etapa de "Planejamento de mini pilotos e iteração" mantem- se ativo o sentimento de ajudar a solucionar possíveis problemas e finalizar recursos pendentes da solução. A etapa de "Criação de um plano de aprendizado" foi considerada a base norteadora da equipe, visto que é primordial estar preparado para possíveis mudanças, melhorias e aprendizados frente a solução que foi proposta inicialmente. Com foco nessa perspectiva de aprendizado, os *softwares* propostos contam com ferramentas para coleta de dados e *logs* de erro. Utilizou-se métricas de validação como o coeficiente de concordância de Kappa, análises de correlação e testes de hipóteses para garantir os acertos dos modelos. Análises qualitativas e quantitativas também vem sendo realizadas pelos especialistas nas soluções propostas. Esses resultados são descritos no Capítulo 7.

Outra validação realizada foi a de acordo com quais os melhores parâmetros a serem utilizados nos modelos. Comparou-se três cenários: (I) As variáveis obtidas na MD; (II) As definidas pelas opiniões dos especialistas; (III) Uma final que leva em consideração as variáveis mineradas com um ajuste feito pelos especialistas.

4.2 CRISP-DM

A usina do estudo de caso possui uma série de dados organizados de forma complexa e semiestruturados. Então, um grande desafio foi o de como obter informações relevantes desse conjunto de dados e o de como identificar variáveis chaves para serem utilizadas como parâmetros nos modelos.

O CRISP-DM é uma metodologia para a MD que vem sendo muito utilizado e alcançando resultados positivos no meio científico. É definido por uma sequência de seis fases que permitem uma compreensão do objetivo da mineração até a extração de informações relevantes para a tomada de decisão.

4.2.1 Compreensão do Problema

Para essa primeira fase do CRISP-DM, já se tinha a definição dos objetivos da pesquisa e problemáticas a serem atacadas. Portanto, o foco desta etapa é definir qual problema a MD deve solucionar, levantando seus requisitos. As atividades gerais dessa etapa são:

(I) Entender o problema de pesquisa. (II) Determinar os objetivos de MD. (III) Coletar os dados iniciais.

Os dados utilizados foram os fornecidos pela usina e os objetivos são: (I) Classificar os dados coletados; (II) Identificar informações, correlações e *outliers* relevantes. A principal problemática é a identificação das principais variáveis, temperatura, umidade, velocidade dos ventos, dentro outras, que influenciam na qualidade e totalidade das safras.

4.2.2 Compreensão dos Dados

Nesta fase, foi de grande importância a colaboração dos especialistas da área de ciências agrárias. Em conjunto, foi-se organizando as bases de dados, colunas e atributos, se identificando melhores nomenclaturas, grupos e formatos de dados. Foi uma etapa minuciosa visando ao entendimento da organização e conexões entre as bases de dados.

A usina possui múltiplas bases de dados, referentes a distintas plantações (áreas) do canavial e variáveis mapeadas. Algumas bases estão em Excel (*XLS*), outras em xml e outras em um banco MySQL. Os dados utilizados neste estudo são referentes a um período de 13 anos, de 2004 a 2017. Os anos de 2003 e 2018 foram excluídos por não se ter as bases totalmente definidas. Das bases existentes, quatro foram escolhidas para serem utilizadas no estudo:

- Safras: Base em formato *XLS* que contém os períodos de início e término das safras, dias totais do período de colheita, produção total, produção própria, produção para fornecedores, área colhida, dentre outros fatores.
- Estação metereológica: Base em formato *XML* que contém os períodos em data e hora das análises, realizadas três vezes por dia. Se captura dados relativos a temperatura de bulbo seco e úmido, umidade relativa, pressão atmosférica, direção e velocidade dos ventos e luminosidade.
- Sensores: Dados georreferenciados e precisos relativos a uma pequena parte da plantação onde se consegue monitorar diversos fatores do solo ligados a safra, como temperatura e umidade.
- Agrônomos: Dados diversos coletados a partir de observações do canavial em diferentes fases de crescimento da cana-de-açúcar. Estes dados contêm observações georreferenciadas de comportamentos e características das plantações em diferentes períodos e fases, realizados quinzenalmente. Além de informações de irrigação, adubação, pragas, precipitações e pH do solo.

O conceito base para a elaboração dos modelos desenvolvidos nesse estudo são os dados, logo é de grande importância que se tenha dados significativos para serem utilizados na modelagem. A Usina São José possui uma dimensão de área plantada gigantesca, sendo

boa parte desta não monitorada. Sendo assim, os dados utilizados neste estudo foram relativos a uma área de 150.000 m^2 (500m x 300m), cerca de 15 hectares, próxima a estação meteorológica da usina e representando uma região onde se tem mais conhecimento e quantidade de dados coletados.

Ao final desta fase se tinha conhecimento sobre como as bases de dados estavam estruturadas e de como as organizar no futuro, já sendo possível idealizar consultas e extrações de interesse para a pesquisa.

4.2.3 Preparação dos Dados

Com o conhecimento obtido na fase anterior, foi elaborado um planejamento para a preparação dos dados para a mineração. Tomando como base os objetivos deste processo, algumas colunas foram eliminadas da base de dados, ou ajustadas. Valores em branco e caracteres especiais também foram ajustados.

Parâmetros como produção própria, produção de fornecedor, quantidade de perfilhos, dentre outros, foram eliminados da análise. Já outros índices como o Toneladas de Cana por Hectare (TCH) foram destacados. Com o auxílio da literatura e dos especialistas pode-se identificar as principais variáveis a serem consideradas.

Tendo-se definido a área do canavial a ser utilizada no estudo e variáveis chaves, se montou a base de dados utilizada nesta pesquisa. Foram unidas as principais dimensões das quatro bases apresentadas anteriormente em um arquivo de formato *CSV*. A base final ficou com 20 variáveis e 14.142 registros. A Tabela 6 ilustra as dimensões utilizadas na base.

Tabela 6 – Descrição da base de dados gerada.

Atributo	**Descrição**
Data	Data do registro (dd/mm/aa)
Hora	Hora do registro (hh:mm)
TempSeco	Temperatura de bulbo seco (°C)
TempUmido	Temperatura de bulbo úmido (°C)
Umid	Umidade relativa (UR%)
Direc	Direção do vento (grau - sentido)
Velo	Velocidade do vento (m/s)
Lum	Luminosidade relativa (Cd)
PreAtm	Pressão atmosférica (Pa)
pHSolo	pH do solo (pH)
TempSolo	Temperatura do solo (°C)
UmidSolo	Umidade do solo (h)
Irrig	Irrigação (booleano)
Adub	Adubação (booleano)
Prag	Pragas (booleano)
Precipt	Precipitação (booleano)
Total	Total do cultivo (ton)
TCH	Total por hectare (ton)
Fase	Fase de crescimento
Quali	Índice de qualidade

Com a base definida pode-se identificar e tratar os *missing values*[4]. Não foram muitos, cerca de 116, mas quando identificados eram substituídos pela média de ocorrência do atributo durante o ano em análise.

Visando as técnicas de análise correlacional e identificação de *outliers* que serão empregadas, também se preparou a base para facilitar a leitura dos algoritmos desenvolvidos em Python, deixando os nomes dos atributos mais acessíveis e claros, retirando caracteres especiais, linhas e colunas em branco.

O indicativo de qualidade foi dado pelos especialistas considerando a fase de crescimento a qual a plantação se encontrava e os demais fatores chaves identificados. Por exemplo, se a plantação se encontra em fase de perfilhamento, sabe-se que alguns dos fatores monitorados são decisivos para a determinação do índice de qualidade. A cana-de-açúcar como uma planta C4[5] necessita de altas intensidades luminosas, sendo a luz o fator mais importante para um perfilhamento de boa qualidade, pois a iluminação adequada na base da planta durante esse período ativa gemas vegetativas basais (MANHÃES et al., 2015). Outros fatores também são importantes, como a temperatura e a umidade.

A Figura 24 ilustra as fases de crescimento da cana-de-açúcar e os fatores determinantes para a sua qualidade, já com um peso relativo a esses para cada uma das fases. Os pesos de cada fase vão de 0 a 10, e através de uma equação linear simples para cada fase é possível se estimar um nível de qualidade. Índices resultantes entre 0 e 5 são considerados de qualidade ruim, entre 5 e 7 de qualidade mediana e entre 7 e 10 de boa qualidade. Os pesos foram determinados de acordo com a literatura (MANHÃES et al., 2015) (SEGATO et al., 2006)(CENTEC, 2004) e ajustados pelos especialistas.

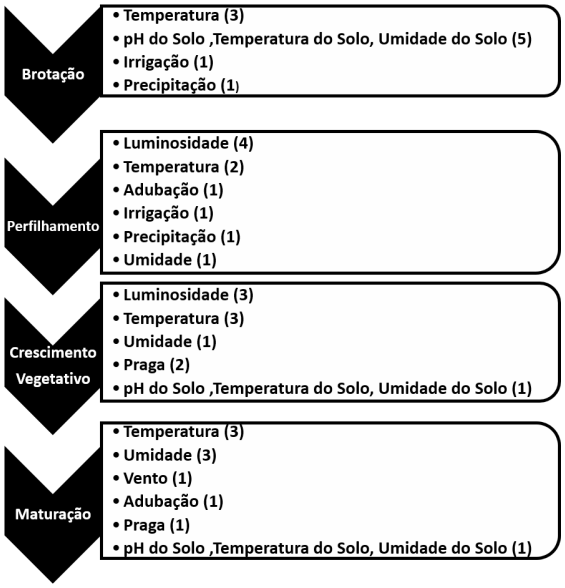


Figura 24 – Pesos dos fatores que afetam as fases de crescimento da cana-de-açúcar

[4] Elementos da base que não possuem nenhum valor associado

[5] Plantas C4 sobrevivem em ambientes áridos, atingindo fotossíntese sob elevada radiação solar.

4.2.4 Modelagem

Na fase de Modelagem deve-se selecionar quais técnicas serão utilizadas no processo de mineração, tendo em vista os objetivos que se deseja alcançar com este processo. Neste trabalho utilizou-se as técnicas descritas na Subseção 3.3.1. Além de análises básicas de identificação de medidas de posição e dispersão, como média, mediana e desvio padrão, utilizou-se análises correlacionais e de identificação de *outliers* nas variáveis de interesse. Todo o processo durou em torno de 2 meses, para se relacionar resultados das análises, especialistas e literatura.

As variáveis de interesse foram identificadas principalmente com o embasamento da literatura e ajuda dos especialistas, então métricas como temperatura, umidade, velocidade dos ventos, luminosidade, dentre outras, foram destacadas nas bases e correlacionadas entre si e principalmente entre o total produzido nas safras.

Com este modelo de análise se tem as variáveis que apresentam maior índice de correlação de influência no ciclo da cana-de-açúcar (Figura 25) em diferentes fases, desde a brotação até a maturação. Essa informação foi muito útil para definição do modelo de cadeia de Markov. Outro dado relevante que se tem é em relação a correlação entre essas variáveis, onde por exemplo, pode-se identificar que com o aumento da umidade tende a haver um aumento na velocidade dos ventos. Esse tipo de análise foi muito útil para a definição do modelo de AC.

A Figura 25 ilustra a matriz de correlação entre os fatores selecionados da base de dados. Quanto mais escura a cor, maior o indicativo de que há uma correlação entre as varáveis. Cores azuladas indicam correlações diretamente proporcionais e cores em vermelho, correlações inversamente proporcionais.

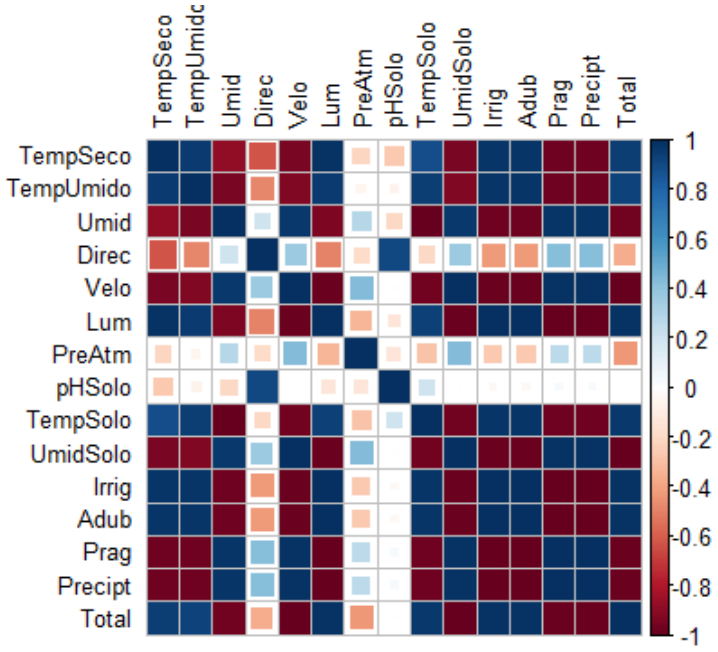


Figura 25 – Matriz de correlação entre os fatores de interesse destacados

Várias observações interessantes e relevantes podem ser realizadas a partir da identificação da matriz de correlação entre os fatores. O primeiro é que a maioria das variáveis destacadas realmente apresentam alto índice de correlação com o total da safra, seja essa direta ou indiretamente proporcional. Variáveis como direção dos ventos, pressão atmosférica e pH do solo, não apresentaram altos índices de correlação. Quanto ao pH, isso possivelmente se deve ao fato da usina controlar bem o pH, tendo esse pouca variação.

Ao observar pragas, por exemplo, pode-se perceber que possuem um alto relacionamento com a umidade relativa e com a velocidade dos ventos. Também é possível observar que durante o período de cultivo, se a média da umidade se mantiver baixa, com ventos em baixa velocidade, temperaturas e lumiinosidade altas, com baixa precipitação, se tem um ambiente favorável para a otimização da safra. Lembrando que em algumas fases é importante que se tenha uma alta umidade, já em outras não, pelo período de maturação ser maior e exigir baixa precipitação, por exemplo, isso influencia no resultado final.

O refinamento final dos fatores estudados visou a identificação das correlações que apresentaram coeficiente (*rho*) maior que 0,3, indicando ao menos um nível moderado de correlação entre os agrupamentos analisados. Outro indicador observado foi em relação ao índice de significância (*p-value*). Por exemplo, para uma relação entre temperatura e safra se tem um coeficiente de correlação alto e positivo de $rho = 0,80$ e um valor significante de $p - value = 0,002353$, logo pode-se supor que há uma relação entre eles.

4.2.5 Avaliação

A avaliação da mineração de dados não foi realizada diretamente com as informações obtidas desse processo. Todas as informações relevantes obtidas da MD foram utilizadas para o desenvolvimento dos modelos de cadeia de Markov e do AC. Esses modelos foram validados por técnicas como a obtenção do coeficiente de concordância de Kappa e testes de hipóteses, verificando se suas simulações estão gerando cenários factíveis.

Quando visto que os resultados não estavam assertivos, os parâmetros foram ajustados por especialistas da área, modificando as probabilidades e indicadores manualmente. Se comparou o desempenho dos modelos propostos utilizando parâmetros extraídos só da mineração de dados, só sugeridos pelos especialistas e utilizando dados da mineração ajustados por especialistas.

4.2.6 Aplicação

Na fase de aplicação o processo definido para a MD e as informações geradas são estruturados e documentados para formalizar uma solução para o problema.

Nesta pesquisa, os resultados do processo de mineração foram aplicados aos modelos estudados, esses foram testados, validados e documentados em conjunto com o processo de mineração e ajustes dos seus parâmetros utilizados.

Capítulo 5

Modelo Cadeia de Markov de Monte Carlo

Realizado o mapeamento sistemático e a aplicação do *Toolkit* HCD, se tem uma base teórica e prática para a definição do modelo MCMC para aplicação as necessidades da agroindústria. Para a definição do modelo aqui descrito, foi seguida a metodologia apresentada na Figura 20. A Figura 26 é uma simplificação da Figura 20 ilustrando apenas os passos utilizados para a definição do modelo aqui descrito.

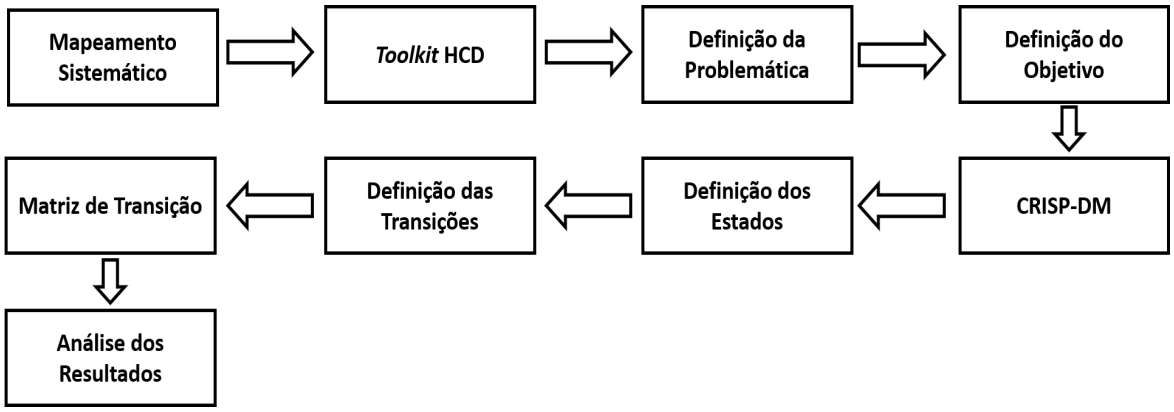


Figura 26 – Metodologia para o desenvolvimento do modelo MCMC

Conforme descrito na Seção 3.5, as abordagens de cadeia de Markov atingem diversas áreas e, em cada área, as soluções são bem particulares. Essa diversidade de aplicação e ampla abrangência dificultam a definição do modelo, mas ao mesmo tempo abrem os horizontes das possibilidades a serem consideradas, permitindo a modelagem de distintos eventos contanto que se tenha conhecimentos detalhados e específicos sobre esses.

Neste capítulo é apresentado como foi desenvolvido o modelo de predição de safras utilizando como base MCMC, discutindo como foram definidos os estados, regras de transição, ciclos temporais e implementação da modelagem.

5.1 Definição dos Estados

Tendo-se definido a problemática e o objetivo ao qual se deseja alcançar, frente aos estudos de caso, é preciso conhecer bem a dinâmica do sistema analisado e o que se deseja resolver. Os estados do modelo são definidos baseando-se em duas características: (I) Propriedades observáveis do sistema; (II) Informações do sistema que se deseja obter.

O sistema abordado neste estudo de caso é relativo a dinâmica do processo de cultivo industrial da cana-de-açúcar, se atentando em todas as fases desde a plantação até a colheita. O grande objetivo que se deseja alcançar é de um modelo capaz de estimar o total produzido por safras.

Logo, observando a dinâmica do processo de cultivo da cana-de-açúcar, embasado na literatura e validado por especialistas, tem-se que as fases de interesse para esse cultivo dentro de um ciclo industrial são: brotação, perfilhamento, crescimento vegetativo e maturação. Sendo assim, essas fases logo foram destacadas como estados fundamentais da cadeia.

Porém, o foco do modelo é a predição de safras, ou seja, predição de colheita total ao final do ciclo produtivo. Então na definição dos estados é essencial a existência de mais dois novos: (I) Um estado de colheita, onde ao final da maturação deve-se apontar para ele e ele que irá referenciar a quantidade total produzida ao final da execução; (II) Um estado de morte, ou improdutividade, que pode ocorrer a qualquer momento e em qualquer estado da cadeia. A Figura 27 ilustra de forma sucinta esse fluxo entre as fases de crescimento da cana-de-açúcar e estados de colheita e morte. O estado de brotação é representado pela letra “B”, perfilhamento pela “P”, crescimento vegetativo pela “C”, maturação pela “M”, colheita pela abreviação “Col” e morte por “Mort”.

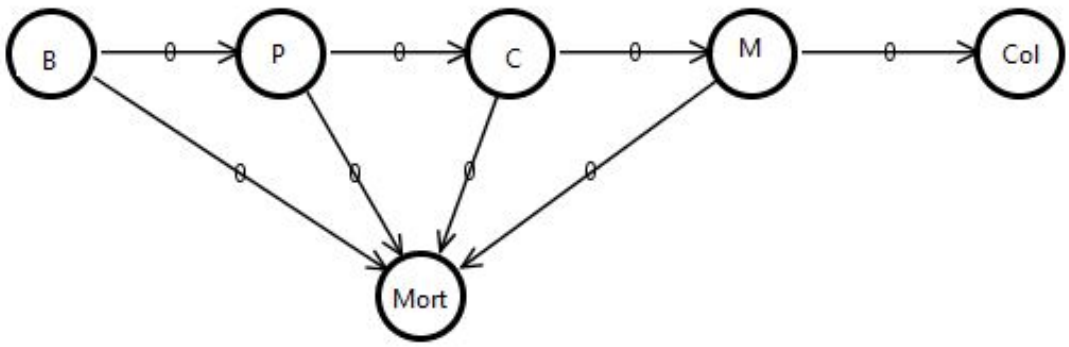


Figura 27 – Estados básicos do modelo

Frente as entrevistas, foi relatada a importância de se prever a qualidade em cada estado produtivo. Logo, os estados de brotação, perfilhamento, crescimento vegetativo e maturação foram divididos cada um em três grupos, os quais são bons, médios e ruins.

Nessa classificação "B1" representa um estado de brotação em boa qualidade, "B2" um estado de brotação de qualidade mediana e "B3" um estado de brotação de má qualidade. A mesma classificação foi feita com os demais estados que representam fases de crescimento e colheita da planta. Esses novas estados são úteis para simulações com uma menor quantidade de ciclos, em que se pode prever além da quantidade a qualidade em que se encontra as plantações. Essa classificação também é muito útil nos últimos estados, podendo prever a qualidade da cana colhida na safra.

A inclusão dos níveis de qualidade dos estados foi realizada frente a uma necessidade solicitada pelos especialistas e a definição desses níveis de qualidade também foram realizadas por eles. Os estados da cadeia são estados booleanos discretos, ou seja, as plantas estarão, ou não, em um estado específico em um ciclo temporal específico.

Os estados foram definidos observando a dinâmica do sistema e a necessidade do que se pretendia analisar, que ao final consistiu no total e qualidade da safra. Os estados foram definidos segundo as características descritas, frente a validações com a literatura e especialistas e por existir base de dados para representar e estimar os seus parâmetros de transição.

5.2 Definição das Transições

Uma vez que os estados foram definidos, é preciso identificar as transições e calcular as suas probabilidades. As transições são identificadas observando-se a sequência temporal dos estados e classificando as transições lógicas possíveis. Basicamente existem dois tipos de transições possíveis, o sistema pode permanecer no mesmo estado, ou pode mudar de um estado para outro.

Para identificar e classificar as transições possíveis foi definido um ciclo temporal para o modelo, em que cada novo ciclo representa um período de quinze dias. Esse período em específico foi escolhido porque a base de dados contém atualizações dos dados a cada quinzena, e também foi um período observacional classificado como eficiente pelos especialistas.

Posteriormente foi observado o período médio de cada fase de crescimento da cana-de-açúcar. Como referencial do ciclo de "cana de ano e meio" (ciclo mais utilizado para obter altas produtividades no primeiro ano), a fase de brotação dura em torno de 30 dias, a fase de perfilhamento 120 dias, a fase de crescimento vegetativo 120 dias e um período de maturação de 210 dias.

Além da transição entre as fases, foi notado a transição entre estados de qualidade entre essas fases. Foi observado que a cada 15 dias a plantação pode transitar porcentagens de seus cultivos entre uma maturação de qualidade boa, ou mediana, por exemplo.

Tomando como base as informações de ciclos de 15 dias, os períodos de cada fase de crescimento da cana-de-açúcar e os estados de qualidade mapeados, pode-se definir

todas as transições possíveis do modelo. Exemplificando, o modelo sempre se inicia por um ou mais de um estado de qualidade da brotação (B1, B2 e/ou B3) e após um ciclo pode transitar entre esses estados de qualidade e/ou morrer. No segundo ciclo, quando completado o período de 30 dias, toda a produção passa para um dos estados de qualidade da fase de perfilhamento (P1, P2 e/ou P3) e/ou morre. Essas transições seguem até chegar na fase de maturação, em que após os últimos 210 dias pode-se prever a quantidade da safra classificada por qualidade (Col1, Col2 e/ou Col3). A Figura 28 ilustra as transições possíveis.

É importante ressaltar que na Figura 28 não estão sendo exibidas as transições para o próprio estado de origem, evitando que o entendimento da imagem se torne complexo, mas os estados B1, B2, B3, C1, C2, C3, P1, P2, P3, M1, M2 e M3 possuem transições para si próprios.

Identificadas as transições, o próximo passo é quantificá-las. Nesta etapa as transições são quantificadas através de suas frequências relativas e probabilidades, respeitando a propriedade de matrizes estocásticas. Para o cálculo e identificação destas, se levou em consideração a sequência de estados possíveis, ciclos temporais pré-definidos, análise dos dados extraídos após a mineração com o CRISP-DM e ajustes dos especialistas.

Os dados brutos relativos ao sistema analisado foram tratados durante a mineração. São retirados os dados fora dos padrões de análise ou incompletos, os quais não contribuem para o modelo a ser construído. Estes dados são dados quantitativos, referentes à frequência relativa entre as fases de crescimento, morte, qualidade e safra.

O método escolhido para analisar a quantidade total produzida, por estado da cadeia, foi considerando as suas variações percentuais entre os ciclos de 15 dias analisados. A variação da quantidade de cana-de-açúcar por estado foi calculada pela Equação 29.

$$\%Transição = \left(\frac{q_{i+1} - q_i}{q_i}\right) * 100 \tag{29}$$

Onde q_i é a quantidade total atual de cana-de-açúcar no estado i e q_{i+1} é a quantidade total de cana no estado $i + 1$ seguinte.

Essa verificação foi realizada para todas as transições mapeadas. A Figura 29 ilustra a matriz de transição base desenvolvida para esse modelo. Exemplificando como o cálculo é realizado, dado que q_i é 100, refletindo a quantidade atual da produção no estado i, que pode ser o estado de brotação, por exemplo, e que o estado q_{i+1}, que pode ser o estado de morte, é 10. Tem-se um resultado de -90%, logo pode-se estimar que 10% da quantidade total existente no estado de brotação após um ciclo passa para o estado de morte.

Frente aos valores obtidos, os especialistas realizaram alguns ajustes, visto que segundo eles alguns resultados não faziam sentido e possivelmente estavam refletindo inconsistências das bases de dados. É importante ressaltar que o fator do MMC entra exatamente neste momento. Com os valores da matriz de transição base definidos, é idealizado como será aplicada as iterações do MMC.

Como visto anteriormente, o MMC é um processo iterativo, baseado em um passeio aleatório de probabilidades, onde a solução final é o resultado da média das n iterações realizadas. Para o modelo aqui apresentado, utiliza-se o MMC considerando-se em qual tempo t e em qual estado i se está. Dados esses parâmetros, utiliza-se uma escolha aleatória de um *range* de probabilidades predefinidas. O Algoritmo 1 ilustra esse passo a passo.

A utilização e ajustes no MMC foram necessários dado a dinâmica do sistema observado. Pode-se perceber que no estado de brotação, por exemplo, não faz sentido que haja uma probabilidade da plantação continuar neste estado após 100 dias, logo, após um período de 2 ciclos, ou 30 dias, toda produção que se encontra nos estados de brotação passa para os estados de qualidade de perfilhamento, ou morrem. Com a utilização do MMC e considerando-se os parâmetros de tempo (t) e estado (i) atuais, é possível alterar as probabilidades de transição.

Logo, no exemplo citado anteriormente, no tempo t_1 (se passou 15 dias) em um estado qualquer de qualidade de brotação, tem-se que a probabilidade de transição para qualquer estado de perfilhamento é de 0%, mas há probabilidades para se transitar entre os estados de qualidade de brotação, ou morte. No tempo t_2 (se passou 30 dias), todas as probabilidades de transição para se permanecer nos estados de brotação são zeradas, sendo possível apenas transitar para um dos estados de qualidade do perfilhamento, ou ir para o estado de morte. Este procedimento é repetido até que se chegue ao último estado, após 32 ciclos de tempo e onde se pode prever o total e qualidade da safra. Por utilizar o método de Monte Carlo esses 32 ciclos são repetidos por n vezes, variando as probabilidades de transição dentro do *range* definido, e o resultado final é a média dos resultados obtidos nas n iterações.

5.3 Matriz e Diagrama de Transições

Este passo se refere à organização das transições quantificadas e estados descritos nas seções anteriores. A quantificação das matrizes de transição foi realizada através das variações percentuais observadas, estratégia bastante utilizada quando trata-se de modelagem de cadeias de Markov.

As cadeias de Markov podem possuir uma série de probabilidades de transição e estados que são descritos através de uma matriz de transição. O diagrama de transição permite ter uma visualização mais facilitada da matriz de transição, funcionando como uma representação gráfica de toda a cadeia. Neste diagrama são visualizados os estados, representados por círculos, e as probabilidades de transição entre os estados. Como já mencionado, a Figura 28 mostra uma representação do diagrama de transição do modelo aqui representado. A Figura 28 não ilustra as transições que voltam para o mesmo estado de saída e de baixa probabilidade de transição, para facilitar a visualização e entendimento

dos fluxos básicos e do diagrama como um todo. Logo, as probabilidades de transição são exibidas parcialmente, uma visão detalhada da matriz base de transição pode ser visto na Figura 29.

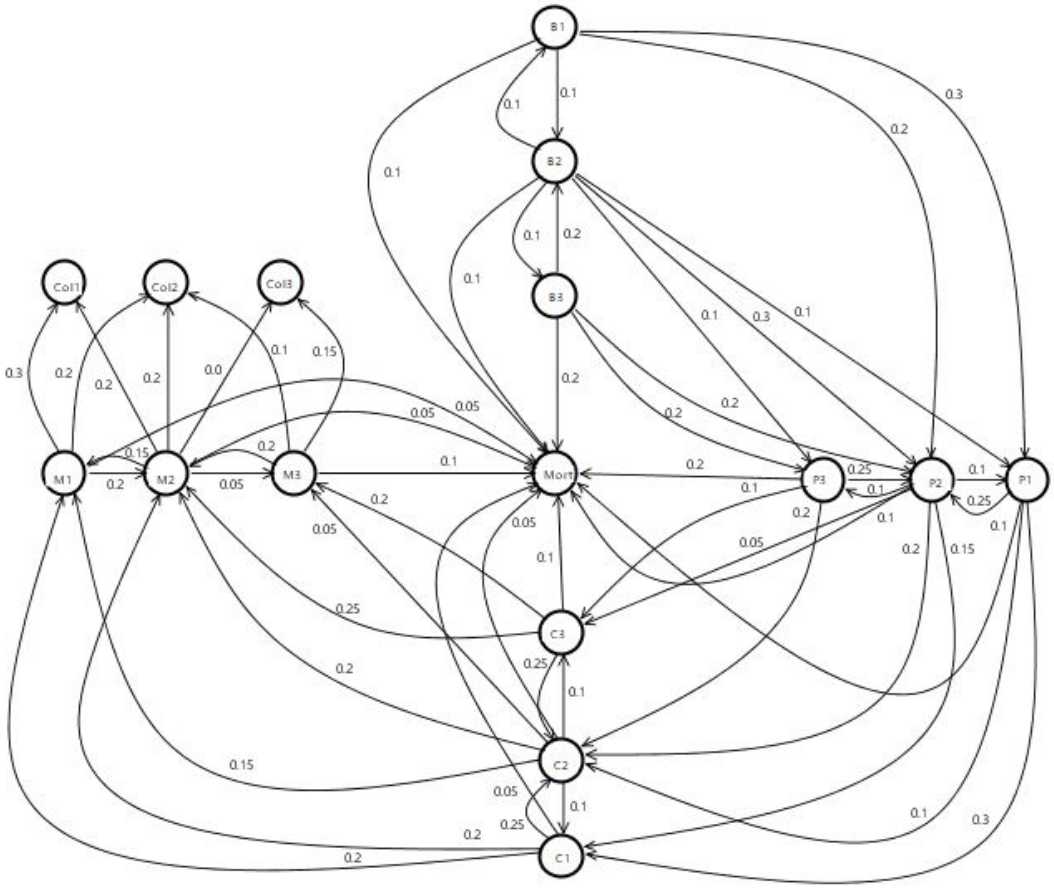


Figura 28 – Diagrama simplificado de transições do modelo

As probabilidades de transição podem ser observadas na Figura 29. Observando as linhas e colunas da matriz de transição pode-se identificar as probabilidades de transição de um estado i qualquer para um $i+1$ qualquer. As linhas representando o estado de partida e as colunas o estado destino.

$$P = \begin{bmatrix}
 & B1 & B2 & B3 & P1 & P2 & P3 & C1 & C2 & C3 & M1 & M2 & M3 & Col1 & Col2 & Col3 & Mort \\
B1 & 0.2 & 0.1 & 0.05 & 0.3 & 0.2 & 0.05 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 \\
B2 & 0.1 & 0.2 & 0.1 & 0.1 & 0.3 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 \\
B3 & 0.05 & 0.2 & 0.1 & 0.05 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 \\
P1 & 0 & 0 & 0 & 0.15 & 0.25 & 0.05 & 0.3 & 0.1 & 0.05 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 \\
P2 & 0 & 0 & 0 & 0.1 & 0.3 & 0.1 & 0.15 & 0.2 & 0.05 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 \\
P3 & 0 & 0 & 0 & 0.05 & 0.25 & 0.15 & 0.05 & 0.2 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 \\
C1 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.25 & 0.05 & 0.2 & 0.2 & 0.05 & 0 & 0 & 0 & 0.05 \\
C2 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.35 & 0.1 & 0.15 & 0.2 & 0.05 & 0 & 0 & 0 & 0.05 \\
C3 & 0 & 0 & 0 & 0 & 0 & 0 & 0.05 & 0.25 & 0.1 & 0.05 & 0.25 & 0.2 & 0 & 0 & 0 & 0.1 \\
M1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.2 & 0.05 & 0.3 & 0.2 & 0 & 0.05 \\
M2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.15 & 0.35 & 0.05 & 0.2 & 0.2 & 0 & 0.05 \\
M3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.05 & 0.2 & 0.35 & 0.05 & 0.1 & 0.15 & 0.1
\end{bmatrix}$$

Figura 29 – Matriz de transição com probabilidades bases do modelo

Como o modelo apresentado utiliza o MMC, os parâmetros ilustrados na Figura 29 não são utilizados de forma determinística. Eles servem como base para o *range* estocástico utilizado nas simulações. Dado que se esteja no estado B1, e no primeiro ciclo, as probabilidades para se passar para o estado P1 são zeradas, por exemplo. Outra situação possível é a determinação da variação das probabilidades a cada iteração de Monte Carlo. Pode-se definir que cada estado pode flutuar suas probabilidades dentro de um *range* de 15%, sendo assim, essa margem de 15% será utilizada para aleatoriamente mudar os valores das transições a cada iteração. A utilização do MMC permite uma aproximação mais exata após uma série de iterações dentro de um *range* possível de probabilidades, a média das safras encontradas após todas as iterações é o resultado final.

5.4 Desenvolvimento do Algoritmo

O modelo foi desenvolvido em Python utilizando as bibliotecas *numpy*, *matplotlib*, *seaborn*, *pylab* e *scipy*. O Python é uma linguagem de programação de alto nível que vem sendo muito utilizada no meio científico, juntamente com o R, para análises e processamentos estatísticos de dados.

Outro motivo para ter se optado pela utilização do Python é o fato de ser uma linguagem simples de ser compreendida e adaptada para outros cenários. Além disso, o Python funciona muito bem como uma linguagem de *back-end*, podendo ser facilmente integrada a plataforma em nuvem também proposta neste trabalho.

O modelo aqui apresentado utiliza como conceito base MCMC, logo o algoritmo desenvolvido foi embasado no Metropolis-Hastings (CHIB; GREENBERG, 1995), que é um método para se obter uma sequência de amostras aleatórias a partir de uma distribuição de probabilidade, sendo já bem definido e validado. O Algoritmo 1 ilustra um pseudocódigo do processamento realizado pelo modelo.

No Algoritmo 1 da linha 1 a 5 são definidos os *inputs* necessários para a execução do modelo, que são os estado de inicialização da colheita, o número de ciclos temporais, a matriz de transição base, o número de iterações de Monte Carlo e o modelo de estado base. As linhas 6 e 7 demonstram os *outputs* esperados após a execução do modelo, logo a safra total e a safra total por qualidade. Na linha 9 se carrega o estado inicial da plantação, no modelo de estados base. As linhas 10 e 11 são responsáveis pelos ciclos de Monte Carlo e temporal. Na linha 12 se utiliza o passeio aleatório dentro de um *range* na matriz de transição base, com isso a matriz é calculada na linha 13. Na linha 14 verifica-se a viabilidade do tempo e da fase de crescimento atuais, caso sejam viáveis, na linha 16 é executada a simulação do próximo estado levando em consideração o modelo de estados base e a matriz de transição base. Todos os resultados são adicionados a um *array* e ao final da simulação esse é tratado para se ter a safra total e a safra total por qualidade.

Algoritmo 1 Modeling and Simulation with Monte Carlo Markov Chains

1: **Input:** $initStateCrop$
2: **Input:** $numberTimeCycles$
3: **Input:** $baseTransitionMatrix$
4: **Input:** $numberMonteCarloIterations$
5: **Input:** $baseModelStates$
6: **Output:** $totalCrop$
7: **Output:** $totalCropForQuality$
8: *Initialize*
9: **Load** $initStateCrop$ **in** $baseModelStates$
10: **For** $k = 1$ to $numberMonteCarloIterations$ **do**:
11: **For** $t = 1$ to $numberTimeCycles$ **do**:
12: **Use** $monteCarloRandomRange$ in $baseTransitionMatrix$
13: **Calculate** $baseTransitionMatrix_t$
14: **Verify** $time_t$ and $currentPhaseSugarCane$ in $baseModelStates$
15: **If** $Verify$ **do**:
16: **Simulate** nextState ($baseModelStates$, $baseTransitionMatrix_t$)
17: **End if**
18: **End for**
19: $resultArray$.insert($resultSimulate_k$)
20: **End for**
21: **Return** $sum(resultArrayAverage)$, $resultArrayAverage$

O modelo descrito neste capítulo é genérico, podendo ser aplicado e testado em diferentes plantações de cana-de-açúcar, sendo possível ajustar as probabilidades de transição. O modelo foi validado em um estudo de caso em uma usina sucroenergética. O Capítulo 7 descreve este estudo, na Seção 7.2 é descrita a utilização e validação deste modelo para a predição da quantidade e qualidade produzida nas safras de cana-de-açúcar da usina.

Capítulo 6

Modelo Autômato Celular Estocástico

O modelo para a predição do total e qualidade da produção de safras apresentado no capítulo anterior é muito útil para a usina, bem como para os agrônomos e pesquisadores dela. Porém, foi observada a necessidade de um modelo que pudesse suportar uma visualização georreferenciada e em tempo real do atual cenário do canavial, como também que fosse possível utilizar este modelo para realizar simulações de cenários futuros de forma aleatória e/ou guiada. Daí surgiu a ideia de se desenvolver um modelo embasado em AC.

Semelhante ao que aconteceu com o modelo de MCMC, o desenvolvimento do modelo de ACE só foi possível graças ao mapeamento sistemático realizado e a aplicação do *Toolkit* HCD. Com os dados e informações obtidos da MD, da literatura e das entrevistas, pode-se traçar um percurso metodológico para o desenvolvimento deste modelo, como ilustrado na Figura 20. A Figura 30 é uma simplificação da Figura 20 ilustrando apenas os passos utilizados para a definição do modelo aqui descrito.

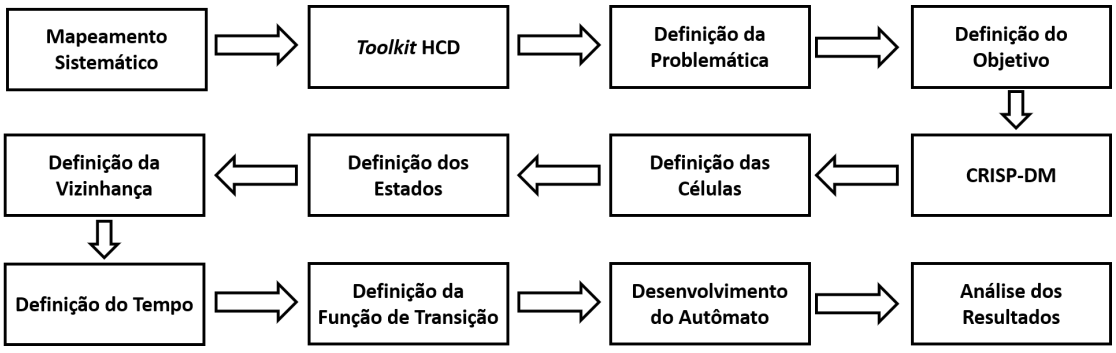


Figura 30 – Metodologia para o desenvolvimento do modelo ACE

Modelos baseados em AC são adequados para refletir características gerais e complexas da realidade, possibilitando uma visualização para tomadas de decisão relevantes. Eles permitem simular a estrutura global de sistemas por meio da aplicação de regras básicas de comportamento de pequena escala dos diversos elementos individuais.

Neste capítulo é apresentado como foi desenvolvido o modelo de monitoramento e predição de cenários das plantações utilizando como base ACE, discutindo como foram

definidas as células, estados, vizinhanças, regras de transição e ciclo temporal.

6.1 Descrição do Modelo

Como visto na Seção 3.6, pode-se dizer que um AC basicamente é constituído de uma matriz, ou grade, que contém células. A evolução do modelo se dá por passos discretos de tempo. Cada célula é caracterizada por um estado pertencente a um conjunto finito de estados. Cada célula evolui de acordo com as mesmas regras que dependem somente do estado em que a célula se encontra e de um número finito de vizinhos. Por fim, a relação com a vizinhança é local e uniforme. Disso, pode-se resumir que para a definição de um AC é preciso classificar suas células, estados, vizinhanças, regras de transição e ciclos temporais.

O AC aqui descrito foi definido no domínio bidimensional, com vizinhança de Moore e dinâmica definida por uma função de transição probabilística. A matriz do autômato representa uma área cultivada, ou um talhão, de uma plantação específica de cana-de-açúcar e, a fim de adequá-la, as colunas correspondem às linhas da área cultivada. Dentro de cada coluna da matriz, encontram-se um conjunto de plantas de cana-de-açúcar dispostas em linhas. Coerentemente, cada célula da matriz representará um conjunto de plantas de cana-de-açúcar que possuírem um valor de estado associado a elas.

Com os resultados das entrevistas, durante a fase de imersão do *toolkit*, pode-se perceber que a representação bidimensional facilita os agrônomos a realizar as simulações, acompanhar o estado atual em tempo real e inserir novos dados ao modelo, principalmente quando utilizado de *smartphones* ou *tablets*. A Figura 31 ilustra o modelo de grade utilizado para essa modelagem.

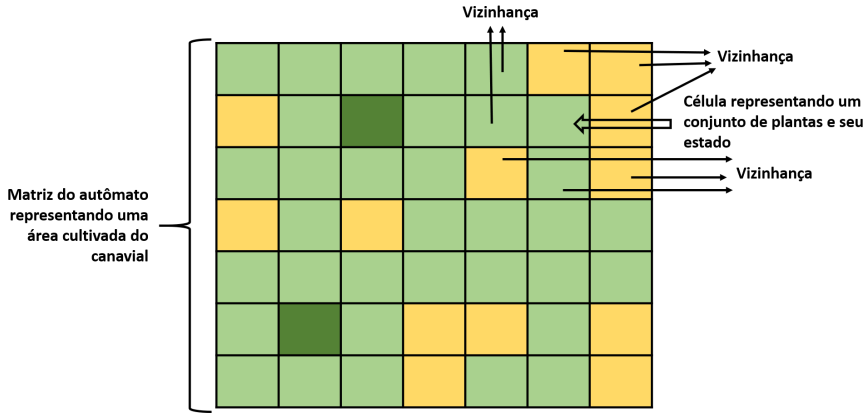


Figura 31 – Matriz do autômato representando uma área cultivada do canavial

Como mostrado no Capítulo 4, a Usina São José cedeu uma área de 150.000 m^2 (500m x 300m), cerca de 15 hectares, para a realização da pesquisa. Essa área foi dividida em subsetores que representam as células do autômato. Frente ao cruzamento das in-

formações obtidas como resultado do processo de MD e ao *feedback* dos especialistas, pode-se perceber que é possível criar agrupamentos na plantação a cada 25 m^2, onde estes possuem o mesmo comportamento relativo as dimensões mapeadas de temperatura, umidade, dentre outras variáveis analisadas. Além de representarem um setor específico, esses agrupamentos conseguem interferir em sua vizinhança. Entende-se que cada célula da matriz representa um conjunto de plantas de cana-de-açúcar que estão na mesma região e possuem um valor de estado associado a elas.

Os estados considerados para as células nesse modelo são semelhantes aos existentes no modelo de MCMC. Independente da fase de crescimento a qual a plantação de cana-de-açúcar se encontra, pode-se classificá-la como sendo de qualidade boa, média, má, morta ou colhida. Para a representação destes estados no AC se atribuiu uma cor para cada um deles, como ilustrado na Figura 32.

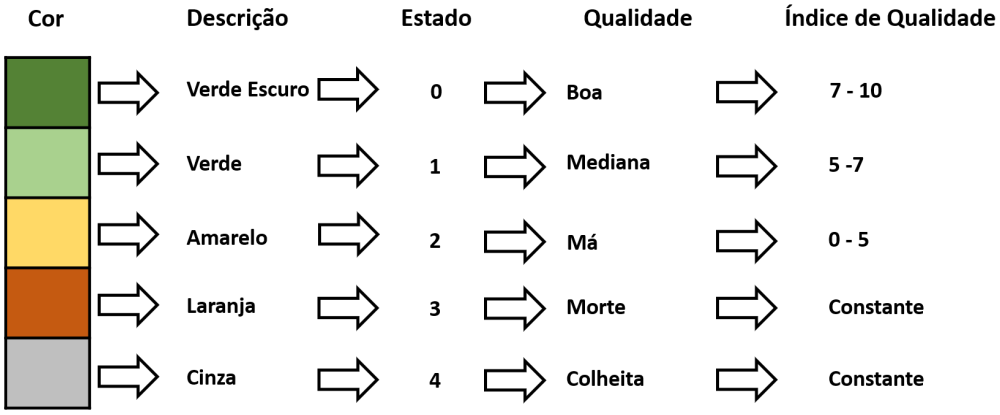


Figura 32 – Representação dos estados possível para as células

Os índices de qualidade são definidos de acordo com o que foi apresentado na Figura 24 no Capítulo 4. Nas simulações, para prospecção de variáveis e fatores futuros, é realizado o cálculo da média destes fatores em relação ao período equivalente em anos anteriores. Por exemplo, o modelo defini a temperatura em setembro de 2020 utilizando a média das temperaturas observadas entre todos os meses de 2020 existentes na base de dados, removendo possíveis *outliers*.

Com as células e os estados possíveis definidos, pode-se começar a elaborar uma regra (ou função) de transição para o autômato. As regras de transição compõem o elemento mais importante para que a aplicação do modelo com AC resulte em simulações confiáveis. Muitas são as abordagens para a geração das regras, e neste modelo é utilizada uma função de transição probabilística baseada em múltiplos critérios.

A construção da função de transição foi embasada nas informações obtidas como resultado da etapa de MD e foi ajustada frente ao que foi proposto pelos especialistas. São considerados aspectos atuais da célula e dos seus vizinhos para determinar o próximo estado. O objetivo é acompanhar o cultivo, podendo estimar cenários futuros da

plantação, no qual é possível estimar se uma determinada região afetará negativamente a produtividade de seus vizinhos.

Uma célula x tem o seu estado alterado na próxima iteração dependendo do estado das células vizinhas. O próximo estado da célula $c(i, j)$, onde i é a linha e j é a coluna, denominado como $S'(c(i, j))$ depende do estado atual de $c(i, j)$, denominado $S(c(i, j))$, e dos estados de seus vizinhos, em uma vizinhança de tamanho 8 (Moore). Além da possibilidade de uma célula de má qualidade piorar o estado da célula atual, ele também pode melhorar, ou pelo menos amenizar o dano. Cada vizinho afeta a célula $c(i, j)$ de forma ponderada, segundo a classificação definida pela literatura e ajustada por especialistas e pelo resultado da fase de MD, ilustrada na Tabela 7. A influência ponderada de cada vizinho se baseia na quantidade de vizinhos de uma determinada qualidade, da média aritmética de qualidade da vizinhança (Equação 30), do estado atual S da célula $c(i, j)$ e da atual fase de crescimento da cana-de-açúcar.

A equação 30 define um $\bar{pv}$, que se trata da média de qualidade da vizinhança ao redor de uma determinada célula. Especificando-se a célula a ser analisada, calcula-se o $\bar{pv}$ levando em consideração a qualidade dos oito vizinhos da célula e com isso se tem uma média de qualidade da vizinhança. O $\bar{pv}$ é utilizado na função de transição e para a determinação do fator de propagação (fp) a ser utilizado. O P indica a probabilidade e o v_n o vizinho que está sendo analisado.

$$\bar{pv} = \frac{\sum_{n=1}^{8} P(v_n(c(i, j)))}{8} \tag{30}$$

A Tabela 7 ilustra os fatores de propagação utilizados na função de transição. Para a determinação de um fator de propagação é preciso se ter a fase de crescimento a qual a plantação se encontra, o estado atual da célula analisada e o $\bar{pv}$ de sua vizinhança. Por exemplo, considerando-se que a plantação esteja em fase de perfilhamento e a célula analisa tem índice de qualidade 6 (mediana) e sua vizinhança possui um $\bar{pv} = 5$, então pode-se dizer que o fp será igual a 0.7. Tanto o fp como o $\bar{pv}$ têm relação direta com a influência da vizinhança sobre a célula analisada no próximo *step* de tempo. As colunas dos estados de colheita e morte não foram adicionadas a tabela por possuírem fator de propagação nulo.

Caso uma forma de combate seja adicionada a plantação (irrigação ou adubo), ou até mesmo algo negativo, como uma praga, denomina-se o fator E na função de transição. Se E for algo negativo, assume um valor que irá diminuir percentualmente a qualidade do estado, caso contrário, assume um valor que irá agregar ao índice de qualidade. O somatório de ações positivas e negativas resulta em um valor utilizado na função de transição. O valor de E é aberto e configurável pelos usuários, podendo os agrônomos indicarem o que acharem ideal caso não seja informado, ou é estimado pelo modelo caso haja um histórico, ou não é considerado no cálculo.

Tabela 7 – Fatores de propagação.

		S(c(i,j))		
		Boa (7-10)	Mediana (5-7)	Má (0-5)
Brotação	$pv \geq 7$	1.25	1.25	1.2
	$7 > pv \geq 5$	0.7	0.6	0.4
	$5 > pv \geq 0$	0.5	0.3	0.1
Perfilhamento	$pv \geq 7$	1.3	1.25	1.25
	$7 > pv \geq 5$	0.8	0.7	0.4
	$5 > pv \geq 0$	0.5	0.3	0.1
Crescimento Vegetativo	$pv \geq 7$	1.2	1.2	1.1
	$7 > pv \geq 5$	1.1	1.1	1.1
	$5 > pv \geq 0$	0.6	0.5	0.3
Maturação	$pv \geq 7$	1.2	1.2	1.2
	$7 > pv \geq 5$	1.1	1.1	1.1
	$5 > pv \geq 0$	0.7	0.7	0.5

As regras que controlam a dinâmica da simulação, alterando os estados das células de modo a representar os cenários da plantação na matriz são definidas por um conjunto de parâmetros, como vem sendo aqui descrito. Os ajustes realizados pelos especialistas nos parâmetros foram importantes para aprimorar e aproximar o resultado da simulação com a realidade da plantação estudada. A dinâmica do autômato pode ser resumida pela Equação 31 que define a função de transição do autômato.

$$S'(c(i,j)) = \frac{S(c(i,j)) + \bar{pv} + \sum_{k=0}^{n} E_k * S(c(i,j)) * fp}{3} \tag{31}$$

A operação é dividida por três para normalizar o valor obtido, também por esse motivo, se o valor resultante for abaixo de zero, é considerado como valor final o zero, se o valor resultante for acima de 10, é considerado que o valor final é 10. Essa normalização é realizada para manter os índices de qualidade entre 0 e 10. Células com $S(c(i,j)) \leq 1.5$ são dadas como mortas. O estado de Colheita é resultado dos ciclos temporais necessários até se chegar a este estado.

A função local de transição é aplicada simultaneamente em todas as células. O estado de uma célula no tempo $t+1$ depende do estado dela própria e dos seus vizinhos no estado t. A vizinhança define o estado das células no próximo instante de tempo. As regras locais de transição ficam responsáveis por atualizar o valor de cada célula da rede, com base nos valores das células que compõem a vizinhança do local, de acordo com o tipo de vizinhança e as condições de fronteira. Ou seja, uma célula que apresenta condições ruins de umidade e temperatura, para uma determinada fase de crescimento da cana-de-açúcar, com o passar do tempo afetará a qualidade das células vizinhas.

Quanto aos ciclos temporais, o tempo é representado pelo número de interações do AC sobre o mapa de origem, sendo que o resultado de uma interação serve de mapa-base para a próxima interação de forma cumulativa. O plano de informação resultante de todas as interações representa as mudanças acumuladas ao longo do período considerado. Para a definição de cada ciclo do AC também se utilizou a mesma estimativa aplicada no modelo

MCMC. Cada ciclo representa um período de quinze dias onde a regra de transição irá agir sobre as células alvo e suas células vizinhas.

Pelo fato de o estado corrente do AC ser o único responsável pelo próximo estado global, pode-se dizer que ele segue um processo markoviano. O modelo desenvolvido nesta pesquisa é baseado em uma implementação de ACE, ou seja, os estados das novas entidades são escolhidos de acordo com algumas distribuições de probabilidade dinâmica, em um tempo discreto. Isso acrescenta um contexto de aleatoriedade importante para as simulações, pois possibilita a visualização de diferentes cenários possíveis para a toma de decisão. Logo, por se tratar de um ACE é possível durante a simulação se variar entre vários cenários, dentro de *range* viável, de fp e $\bar{pv}$, por exemplo, possibilitando achar uma média mais assertiva entre o total de simulações, semelhante ao realizado com a aplicação do MMC nas cadeias de Markov.

6.2 Desenvolvimento do Algoritmo

O autômato desenvolvido pode ser inicializado por dados fornecidos pelo usuário do *software*, pela análise de um arquivo *json*, ou pela análise em tempo real provinda de um *socket* conectado a um *back-end* que processa os dados oriundos dos sensores. Logo, apenas com esses *inputs* pode-se ter uma visualização em tempo real da plantação, podendo monitorá-la remotamente.

Os dados fornecidos pontualmente também podem ser utilizados nas simulações, eles são cruzados com os dados históricos e processados de acordo com os ciclos temporais selecionados. A variável temporal controla os *steps* de transição de estado das células do AC. As mudanças de estado dos vizinhos de uma célula qualquer, podem representar mudanças climáticas, por exemplo, que vão influenciar nas células vizinhas.

Na representação visual do AC é possível selecionar qualquer célula do *grid* e recuperar informações da amostra correspondente, inclusive dados de temperatura, umidade, estado atual do setor, fase de crescimento, endereço do *grid*, dentre outros. Ao realizar uma simulação é possível adicionar fatores externos como irrigação, adubação e pragas, sendo esses booleanos. A simulação é interativa e simples, na qual o usuário pode pausar, realizar interferências no ambiente, continuar e reiniciar. Este modelo também foi desenvolvido em Python, em que se utilizou as bibliotecas *numpy*, *scipy*, *pylab*, *time*, *random* e *pygame*. A utilização do Python nesta implementação traz os mesmos benefícios observados na implantação anterior, além de ser uma linguagem muito eficiente, possibilita uma integração fácil com plataformas em nuvem.

É desejável que o modelo com base em AC tenha uma interface para a visualização das grades dos autômatos, e para isso utilizou-se a biblioteca *pygame* para gerar a interface das grades. O fato de se estar utilizando ACE, faz com que o algoritmo realize várias simulações variando aleatoriamente (dentro de um *range*) as probabilidades utilizadas na

função de transição, ao final calculando uma média dos valores obtidos. O Algoritmo 2 ilustra um pseudocódigo do processamento realizado pelo modelo.

Algoritmo 2 Modeling and Simulation with Stochastic Cellular Automata

```
1: Input: gridDimensions
2: Input: initCellConfiguration
3: Input: numberTimeCycles
4: Input: numberSimulations
5: Input: externalArrayFactors (pests, irrigation, fertilizer)
6: Output: futurePlantationScenario
7: Initialize
8: Load baseAutomaton(gridDimensions, initCellConfiguration)
9: For k = 1 to numberSimulations do:
10:     For t = 1 to numberTimeCycles do:
11:         For cel in baseAutomaton do:
12:             Calculate pv
13:             Calculate fp
14:             Calculate E(externalArrayFactors)
15:             Use randomRange(pv, fp)
16:             Use transitionFunction in cel
17:         End for
18:         Generate newGrid_t
19:         Simulate nextAuto(newGrid_t)
20:     End for
21:     resultGrid.insert(resultGrid_k)
22: End for
23: Return resultGridAverage
```

No Algoritmo 2 da linha 1 a 5 são definidos os *inputs* necessários para a execução do modelo, que são o *grid* de dimensões, a configuração inicial das células, o número de ciclos temporais, o número de simulações a serem realizadas e um *array* de fatores externos. A linha 6 demonstra o *output* esperado após a execução do modelo, logo o cenário futuro da plantação. As linhas 9 e 10 são responsáveis pelos ciclos das simulações e temporais. A linha 11 é responsável por percorrer cada célula do autômato. Na linha 12 é calculada a qualidade da vizinhança. Na linha 13 é calculado o fator de propagação. Na linha 14 são calculados os fatores externos. Na linha 15 utiliza-se um passeio aleatório com os valores da qualidade da vizinhança e do fator de propagação. Na 16 a função de transição é aplicada nas células. Por fim, é gerado um novo *grid* considerando as transformações célula a célula até se obter um autômato resultante final.

O modelo descrito neste capítulo, como o modelo da Seção 5, também foi idealizado para funcionar de forma genérica, podendo ser aplicado e testado em diferentes plantações de cana-de-açúcar. Este modelo também foi validado no estudo de caso. O Capítulo 7 descreve este estudo, na Seção 7.3 é descrita a utilização e validação deste modelo para a predição de cenários georreferenciados das plantações de cana-de-açúcar da usina.

Capítulo 7

Estudo de Caso

Os modelos e artefatos propostos neste trabalho foram testados e validados na São José Agroindustrial. A usina possui uma infraestrutura agroindustrial completa, com uma capacidade anual de moagem média de 1,3 milhões de toneladas de cana, produzindo mais de 2,7 milhões de sacos de açúcar, mais de 20 mil m^3 de etanol e comercializando mais de 4 MWh de energia elétrica no período da safra (agosto a janeiro). O mix de produção está dividido em 80% para açúcar e 20% para etanol (JOSÉ, 2018).

O setor agrícola da usina envolve as áreas destinadas ao plantio da cana-de-açúcar, matéria-prima da indústria sucroenergética, e conservação do Parque Ecológico São José e demais áreas de preservação e manutenção ambiental, como por exemplo, os rios e açudes. Com o objetivo de ampliar a sua produção agrícola, mantendo a qualidade já validada de seus produtos, a usina investe em P&D e conta há anos com um convênio com a rede Ridesa, da UFRPE, onde há o Centro Experimental de Carpina, que trabalha com pesquisas de novas espécies.

O uso do *Toolkit* HCD foi fundamental para se chegar em soluções propostas que garantem alinhamento com as necessidades dos usuários, pois estes foram envolvidos em todas as etapas do processo de ideação, criação e desenvolvimento. Alguns dados e resultados obtidos das fases do *Toolkit* são relevantes de serem descritos.

As pessoas que participaram do processo de pesquisa caracterizavam distintas faixas etárias, gêneros e áreas de estudo (agrárias e computação). Foram entrevistadas 30 pessoas, destas, 15 (50%) foram classificadas como sendo do público ideal, 9 (30%) medianos e 6 (20%) não ideais. Dentre os entrevistados, 5 (16.7%) foram classificados como especialistas, todos doutores em suas áreas de estudo e colaboradores fundamentais para alinhamento dos modelos desenvolvidos nesse projeto.

Todos os entrevistados durante a fase Ouvir afirmaram que uma das maiores problemáticas no cultivo da cana-de-açúcar é realizar predições de safras, monitorar e prever cenários futuros das plantações. Segundo alguns entrevistados a cana-de-açúcar cresce como capim, sendo impossível prever o seu comportamento. Devido a esta complexidade, todos afirmaram que seria interessante a existência de uma solução para monitoramento

e simulação de cenários das plantações, sendo possível otimizar recursos para se obter safras mais lucrativas.

Um ponto importante das entrevistas foi como se realiza as análises hoje na usina. A maioria dos entrevistados disseram que fazem suas análises e prospecções via planilhas no Excel ou que simplesmente não fazem análises quanto a safra, mas uma estimativa superficial quanto a períodos passados. Os membros não ideais, os quais são mais conservadores, deram uma informação muito relevante e curiosa, que a predição feita por eles é oriunda de conversas com funcionários antigos da usina: "Os canavieiros mais experientes conseguem através de observações das plantas e do clima estimar a qualidade e quantidade da safra melhor do que qualquer planilha". Infelizmente não foi possível nesse estudo realizar uma comparação entre os acertos dos modelos propostos e de canavieiros experientes.

A utilização de uma solução de *software* que possa ser acessada via dispositivos móveis ou computadores foi muito bem vista pelos membros ideais e medianos, principalmente por conta da comodidade na realização dos monitoramentos e simulação, sendo indicado inclusive que houvesse alarmes para mudanças críticas em setores das plantações. Todos os *feedbacks* recebidos em todas as fases do *Toolkit* HCD foram analisados e considerados para a realização de melhorias nas soluções propostas.

Após todo o percurso realizado com a implantação da metodologia do *Toolkit* HCD, chegou-se a uma primeira versão das soluções propostas. Se tem ao todo quatro artefatos para auxiliar e otimizar o cotidiano de trabalho de agrônomos e pesquisadores da usina: (I) Plataforma para armazenamento e compartilhamento de dados das plantações em nuvem; (II) Aplicativo para monitoramento e inserção de dados georreferenciados em tempo real das plantações; (III) Modelo de MCMC para a realização de simulações de safras; (IV) Modelo de ACE para a realização de simulações de cenários futuros das plantações.

A usina já investe em tecnologias para o melhoramento genético de espécies de cana-de-açúcar e para a melhoria do solo, porém percebeu a importância e necessidade de se investir em tecnologias e inovações para o monitoramento das plantações e predições de cenários e safras. Um melhor conhecimento das áreas cultivadas e do que pode acontecer com essas áreas, da poder para a usina otimizar a sua produção enquanto reduz impactos ambientais. Para isso, é proposto a utilização de sensores nas plantações e recursos de software para coleta, análise, monitoramento e simulação dos dados e cenários de cultivo.

7.1 Plataforma e Aplicativo Sugeridos

Como observado durante a fase Ouvir do *Toolkit* HCD, a principal ferramenta para análise de dados utilizada na usina são planilhas em Excel. Porém, ao se observar a tendência de inovação de grandes indústrias, percebe-se que os dias de análises amplamente baseadas em planilhas estão desaparecendo. Departamentos e empresas inteiras estão

buscando suas próprias iniciativas de análise e *business intelligence* (COSTA, 2017).

Para inovar na coleta e monitoramento de dados de suas plantações, são idealizadas soluções que envolvem o uso de IoT e de dispositivos móveis para a facilitação da coleta dos dados. Os sensores podem monitorar em tempo real temperatura, umidade, luminosidade e velocidade dos ventos e os pesquisadores, via aplicativo em dispositivos móveis, podem inserir informações e dados georreferenciados sobre a plantação. Todos esses dados coletados são armazenados e compartilhados em uma infraestrutura em nuvem que facilita o consumo destes dados por outras ferramentas que podem utilizá-los para análises.

Canzian (1999) afirmava que, alguns campos podem ser bem uniformes, mas outros apresentam variações no tipo de solo, fertilidade e outros fatores que afetam a produção agrícola. Se a variabilidade do campo puder ser medida e registrada, estas informações poderão ser usadas para otimizar as aplicações em cada ponto, sendo este o novo conceito de agricultura de precisão.

De 1999 até hoje a computação evoluiu muito e através de sensores, atuadores e aplicativos é possível monitorar e acompanhar variações de variáveis significativas para o ciclo do plantio. Só que para isso ser possível também é preciso se preocupar com uma infraestrutura que consiga armazenar e prover esses dados de forma estruturada.

Diante da coleta de uma vasta massa de dados, a usina se deparou com alguns problemas: I) o armazenamento estruturado e acessível desses dados; II) a centralização dos dados coletados pelos sensores e pelos pesquisadores; III) a georreferência correlacionada entre dados dos sensores e dos pesquisadores; IV) a existência de um ambiente compartilhado, onde pesquisadores de diferentes áreas da usina e de universidades possam compartilhar dados e informações relativas às plantações, tornando o conhecimento compartilhado e construtivo.

Visando a solucionar esses problemas, foi proposta a infraestrutura ilustrada na Figura 33. A plataforma desenvolvida foi estruturada em uma arquitetura na AWS IoT Core. Foram utilizadas diferentes tecnologias e linguagens de programação para a implementação de toda a ferramenta. Pode-se perceber que a aplicação desenvolvida possibilita a integração tanto com os sensores das plantações, como também com dados lançados pelos pesquisadores, ou importados de outras ferramentas. Após a coleta e o processamento, os dados podem ser disponibilizados tanto via API GET como exportados em formato CSV, ou visualizados na aplicação no AC gerado.

Cruzando os resultados dos questionários e *feedbacks* pode-se perceber que a plataforma trouxe ganhos positivos em relação a todos os aspectos analisados quando comparada a utilização de *datasets* locais. A Figura 34 ilustra esses resultados e pode-se observar que em critérios como o de georreferenciamento, exportação de dados e compartilhamento de plantações, os pesquisadores foram unânimes em atribuir nota máxima para a plataforma. Critérios como facilidade para a realização de pesquisas, disponibilidade dos dados e coleta automatizada, também receberam notas muito boas.

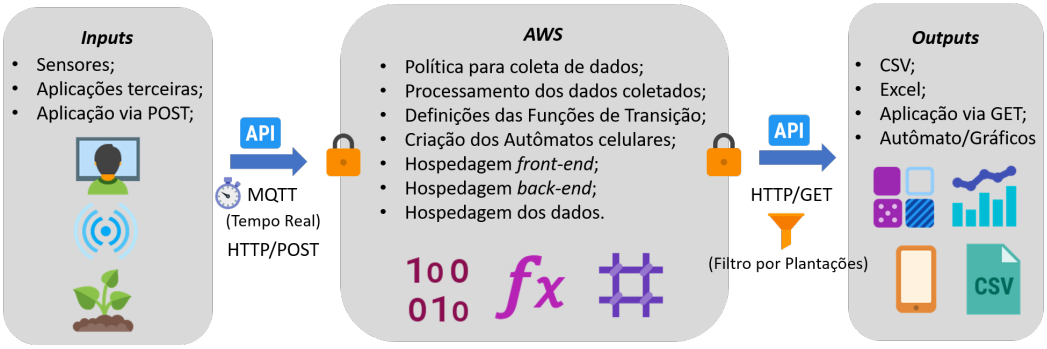


Figura 33 – Arquitetura da Plataforma Desenvolvida

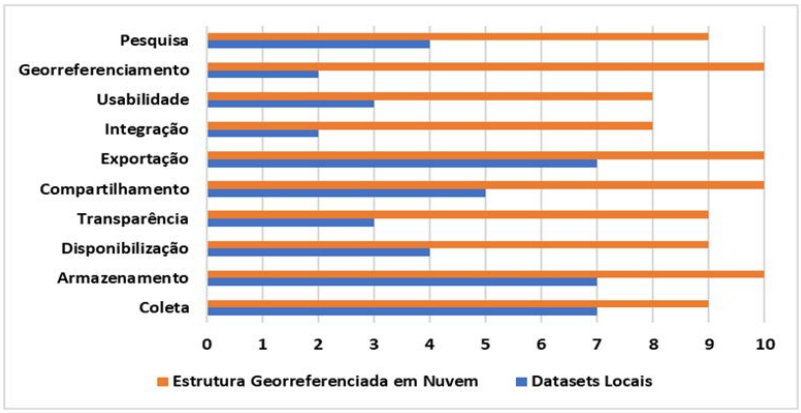


Figura 34 – Comparação entre características de armazenamento e disponibilização dos dados coletados

O aplicativo desenvolvido para monitoramento e inserção de dados da plantação em tempo próximo ao real na plataforma utiliza uma interface de AC de domínio bidimensional, com vizinhança de Moore (oito células ao redor de uma célula central em uma grade quadrada) e dinâmica definida por uma função de transição probabilística. Com os resultados das entrevistas pode-se perceber que a representação bidimensional da grade facilita os agrônomos a realizarem as simulações, acompanhar o estado das plantações em tempo real e inserir novos dados ao modelo, principalmente quando utiliza-se *smartphones* ou *tablets*.

A matriz do AC representa a área cultivada da plantação. Dentro de cada coluna da matriz encontram-se as células dispostas em linhas, que representam um conjunto de plantas de uma região específica da plantação com dimensão adaptável a quantos metros quadrados forem indicados pelos pesquisadores. Entende-se que cada célula da matriz representa um conjunto de plantas que estão na mesma região e possuem um valor de estado associado a elas.

A aplicação desenvolvida possibilita o cadastro de plantações, bem como a visualização dos AC criados para a representação destas. A Figura 35 ilustra a interface de visualização

das plantações em AC na plataforma. Ao clicar em uma célula do autômato pode-se analisar características específicas daquela região plantada.

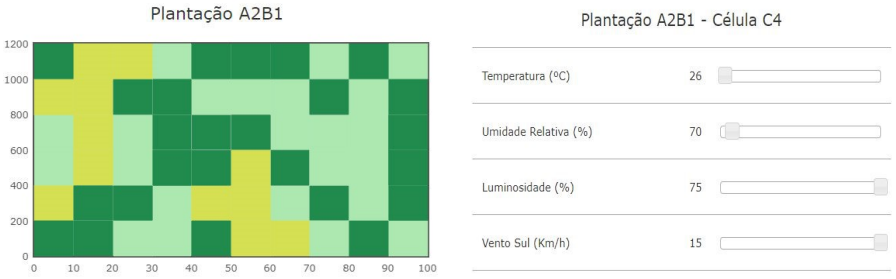


Figura 35 – AC e Detalhamento de uma Célula no aplicativo

O módulo de visualização do AC é mais utilizado para tomadas de decisões estratégicas por parte dos gestores. O módulo mais utilizado pelos pesquisadores são os das APIs disponibilizadas pela plataforma. Com um simples clique pode-se baixar dados e informações de uma plantação específica. Em uma requisição GET a plataforma retorna um objeto json com dados estruturados considerando a plantação e suas células. O uso dessas APIs possibilita a interação em tempo real com a plantação, permitindo aos pesquisadores a utilização de modelos de previsão e a construção de simulações para a otimização da safra. A ferramenta desenvolvida facilitou a colaboração de pesquisas realizadas dentro da usina e em universidades possibilitando a troca de conhecimentos entre os pesquisadores.

Tão importantes quanto os dispositivos usados na agricultura de precisão, perceber que a informação usada ou coletada é o ingrediente chave para o sucesso do sistema. O conceito de agricultura de precisão se distingue da agricultura tradicional por seu nível de manejo. Em vez de administrar uma área inteira como uma única unidade, o manejo é adaptado para pequenas áreas dentro de um campo (TSCHIEDEL; FERREIRA, 2002).

É de grande importância coorporativa e industrial que se tenha uma infraestrutura moderna e atualizada para a coleta, armazenamento e monitoramento de dados críticos da empresa. Porém, embora a maioria das organizações tenha uma forte capacidade de coletar e armazenar dados históricos, elas continuam a ter dificuldades para transformar dados em análises significativas (BORLIDO, 2017). Por esse motivo, o grande foco desta pesquisa é a utilização dos dados coletados e minerados nos modelos de simulação, para que esses deem informações relevantes para a tomada de decisão dos gestores, pesquisadores e agrônomos.

7.2 Simulação e Experimentos com o Modelo Cadeia de Markov de Monte Carlo

O Capítulo 5 descreve o modelo MCMC apresentado neste estudo para a simulação de safras e de seus índices de qualidade. O modelo descrito é genérico, e a ideia é que este

possa ser aplicado a distintas plantações em diferentes regiões. O modelo foi desenvolvido tomando como base dados da usina São José e ajustado de acordo com a literatura e especialistas. Esta seção apresenta os resultados obtidos com a utilização deste modelo em campo.

Como já relatado anteriormente, a área disponibilizada para este estudo foi de 150.000 m^2 (500m x 300m), correspondente a 15 hectares. E é de grande importância para os agrônomos e gestores da usina saberem uma projeção da safra dado a quantidade plantada inicialmente, ou o estado atual da plantação, como também, a qualidade da safra. Para isso o modelo aqui apresentado foi desenvolvido e posteriormente testado para a realidade da usina, visando a viabilidade de uma modelagem computacional através de MCMC para realizar essas estimativas.

Como a área disponibilizada para o estudo foi de 15 hectares, utilizou-se inicialmente como base as métricas médias da usina por hectare, considerando toda a dimensão de área plantada do canavial. A fase de plantio até a colheita da safra, são importantes para a realização das simulações, atuando como métricas de entrada e estimativas dos resultados do modelo. Para isso, utilizou-se a base de dados das plantações nas simulações, comparando os resultados das simulações com os observados nas bases de dados, frente a um cenário ou período comum entre ambos. As validações se deram via a obtenção do coeficiente de correlação, teste de hipótese, concordância de Kappa e validação cruzada. A Tabela 8 contém os dados referentes a Toneladas de Muda por Hectare (TMPH), Toneladas de Cana por Hectare (TCH) e Toneladas de Pol (açúcar) por Hectare (TPH) no período entre 2004 e 2017.

Tabela 8 – Histórico geral de safras em toneladas por hectares.

Período/Anos	**04-05**	**05-06**	**06-07**	**07-08**	**08-09**	**09-10**	**10-11**
TMPH	11,17	10,50	11,50	11,86	12,13	11,12	12,02
TCH	58,12	50,28	62,77	71,02	64,53	63,03	61,54
TPH	8,15	6,95	8,40	9,73	8,91	8,38	8,12
Período/Anos	**11-12**	**12-13**	**13-14**	**14-15**	**15-16**	**16-17**	/
TMPH	10,85	11,20	11,43	11,68	11,51	11,56	/
TCH	63,85	50,48	52,90	64,73	55,73	56,87	/
TPH	8,43	6,94	6,56	8,07	7,25	7,93	/

Em uma análise inicial, correlacionando TMPH com TCH, se obtém um índice de 0.5243407 com um $p-value$= 0.06584. Se correlacionando TMPH com TPH, obtém-se um coeficiente ainda menor, de 0.464036 com um $p-value$=0.1102. Porém, ao se correlacionar TCH com TPH atinge-se um coeficiente de correlação elevado, de 0.9360265, com um $p-value$=2.55e-06. Logo, pode-se supor que a quantidade inicial plantada não possui um alto índice de correlação com a produção final colhida, tendo outros fatores influenciando no aumento ou diminuição da safra. Por outro lado, percebe-se que a quantidade de açúcar produzido está diretamente relacionada a safra final colhida, podendo ser equiparado e identificado com uma maior facilidade.

Sabendo disso, se adicionou ao modelo um *script* simples para prever a produção de açúcar e etanol dado o total estimado de cana colhida. É levado em consideração a média histórica de TCH, TPH e etanol. Sempre que a usina produz açúcar também possui uma quantidade menor de etanol, sendo este um dos resultados possíveis. Outra condição é da produção só de etanol dado o total de cana obtido na safra. As Equações 32 e 33 ilustram esses cálculos.

$$\text{Total Açúcar + Etanol} = \frac{\text{TCH x 130kg}}{1000} + \text{TCH x 14,6L} \tag{32}$$

$$\text{Total Etanol} = \text{TCH x 97,2L} \tag{33}$$

Como pode-se comprovar que o TMPH não possui um alto coeficiente de correlação com TCH ou TPH, aplicou-se o modelo MCMC utilizando como entrada o TMPH. A Tabela 9 ilustra os resultados obtidos pelo modelo em comparação com os existentes na base de dados da usina.

Tabela 9 – Resultados do modelo simulando dados por hectares.

Período/Anos	**04-05**	**05-06**	**06-07**	**07-208**	**08-09**	**09-10**	**10-11**
TMPH	11.17	10.5	11.5	11.86	12.13	11.12	12.02
TCH	58.12	50.28	62.77	71.02	64.53	63.03	61.54
TPH	8.15	6.95	8.4	9.73	8.91	8.38	8.12
TCH-MCMC	57.89	50.33	61.62	70.83	64.47	63.11	61.46
TPH-MCMC	7.5257	6.5429	8.0106	9.2079	8.3811	8.2043	7.9898
TPH(L)-MCMC	845.194	734.818	899.652	1034.118	941.262	921.406	897.316
Etanol(L)-MCMC	5626.908	4892.076	5989.464	6884.676	6266.484	6134.292	5973.912
Período/Anos	**11-12**	**12-13**	**13-14**	**14-15**	**15-16**	**16-17**	/
TMPH	10.85	11.2	11.43	11.68	11.51	11.56	/
TCH	63.85	50.48	52.9	64.73	55.73	56.87	/
TPH	8.43	6.94	6.56	8.07	7.25	7.93	/
TCH-MCMC	64.05	50.52	52.67	64.89	55.41	56.94	/
TPH-MCMC	8.3265	6.5676	6.8471	8.4357	7.2033	7.4022	/
TPH(L)-MCMC	935.13	737.592	768.982	947.394	808.986	831.324	/
Etanol(L)-MCMC	6225.66	4910.544	5119.524	6307.308	5385.852	5534.568	/

Essa simulação utilizou 32 ciclos para se chegar até o período de colheita. Foi desconsiderado os indicadores de total por qualidade, se somando apenas a quantidade total da safra ($Col1 + Col2 + Col3$). Porém, a simulação considerou todos os fatores que influenciam na qualidade e fases de crescimento da planta. Percebe-se que ao correlacionar o TCH da base de dados com o do modelo se tem um coeficiente de correlação de 0.983838 e um $p - value$=1.447e-09. Ao se correlacionar os valores de TPH da base com TPH do modelo, se obtém um coeficiente de 0.9634175 e um $p - value$=1.243e-07. Pode-se perceber que mesmo não se tendo muitos dados dessa base, o modelo apresenta uma boa correlação entre a quantidade total de cana-de-açúcar por hectare e o total de açúcar por hectare. Isso talvez se dê pela relação mais simples que há entre essas variáveis e ao correlacionamento mínimo que já existia entre TMPH e TCH. Através do modelo também

foi possível estimar uma produção de etanol, dado que se escolheu produzir açúcar, como também mostra o cenário da produção única do etanol.

Para todas as amostras se realizou testes de normalidade, onde se verificou o qq-plot e o Shapiro-Wilk para as suas devidas validações. O Shapiro-Wilk para o TCH teve um de $p - value = 0.5614$, já o TPH $p - value = 0.6144$, por exemplo. Se comprovando a normalidade das amostras, foi realizado um teste de hipótese T de Student, comparando os resultados da base com os gerados pelo modelo. O teste de hipótese para o TCH teve um valor t=0.029822 e um $p - value = 0.9765$, indicando que os resultados do modelo coincidem com os da base. Já para o TPH o valor t foi de 0.65714 e o $p - value = 0.5174$. Para ambos os testes *T Student*, as hipóteses consideradas foram:

$$\begin{cases} H_0 & : base = modelo \\ H_1 & : base \neq modelo \end{cases}$$

.

Por fim, realizou-se uma análise do coeficiente de concordância Kappa, buscando identificar o índice de relação entre as amostras. A Tabela 10 ilustra a interpretação dos valores de Kappa e as Equações 34, 35 e 36 mostram a obtenção dele.

Tabela 10 – Interpretação do valor de Kappa.

Valor de Kappa	**Interpretação**
Menor que zero	insignificante
Entre 0 e 0.2	fraca
Entre 0.21 e 0.4	razoável
Entre 0.41 e 0.6	moderada
Entre 0.61 e 0.8	forte
Entre 0.81 e 1	quase perfeita

$$\widehat{K} = \frac{\widehat{p}_0 - \widehat{p}_e}{1 - \widehat{p}_e} \tag{34}$$

$$\widehat{p}_0 = \sum_{i=1}^{r} \frac{n_{ii}}{n} \tag{35}$$

$$\widehat{p}_e = \sum_{i=1}^{r} \frac{n_{i.} * n_{.i}}{n^2} \tag{36}$$

O Kappa para a análise dos TCH foi de 0.72 com um $p - value = 0.00389$, e para o TPH 0.61, $p - value = 0.00337$. Neste teste, para $p - values$ menores que 0.05 pode-se rejeitar a hipótese de que a concordância entre as amostras foi puramente aleatória. Com isso, se obteve índices bons e significativos, comprovando a eficiência do modelo para esse cenário aplicado. Havia poucos dados relativos a simulação por hectare, mesmo assim pode-se obter resultados satisfatórios a nível de correlação e equivalência geral com os dados da base. Isso possivelmente ocorreu graças a utilização do MMC e da análise de fatores de qualidade do modelo.

A Figura 36 ilustra os resultados das 100 simulações realizadas para a determinação do TCH de 2004 a 2005. Esses resultados ilustram a utilização do MMC na identificação do resultado final, melhorando a precisão frente a média de várias simulações e alternâncias nas probabilidades da matriz de transição.

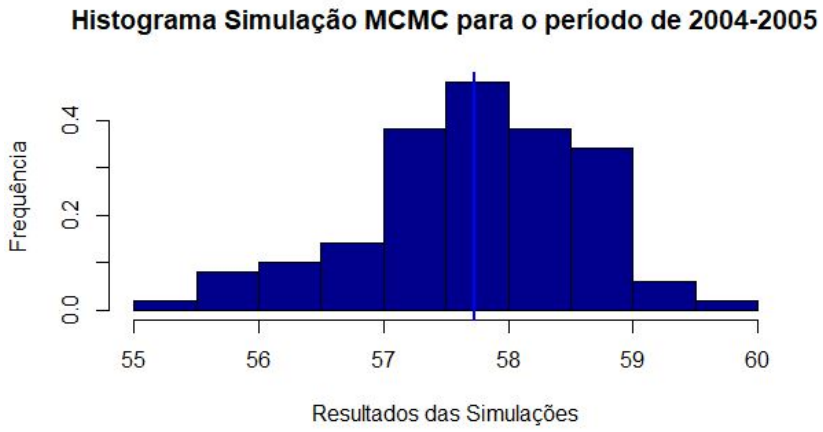


Figura 36 – Simulações realizadas para a determinação do TCH do período de 2004-2005

A grande complexidade da modelagem está em se estimar safras de regiões específicas e considerando seus índices de qualidade. A simulação anterior apenas pegou o somatório dos três estados de colheita, já que na base geral da usina não se considera o fator qualidade por hectare. Para a realização destas novas simulações se utilizou os dados dos 15 hectares de região mapeada, com dados mais refinados.

Semelhante ao experimento realizado anteriormente, porém agora se considerando todos os fatores e resultados, se testou o modelo proposto com os dados e todos os fatores presentes na base de dados. O experimento utilizou a base de mineração com dados de 2004 a 2014, deixando de fora os de 2015 a 2017 para outras validações. Logo, separando um conjunto de dados para o treinamento do modelo e outro para as validações, como indicado pelo processo de validação cruzada. As bases de 2015 a 2017 possuem dados mais recentes, sendo mais fácil o confronto dos resultados do modelo com os existentes na base.

As Figuras 37 e 38 representam, respectivamente, os resultados das simulações para os períodos de 2004 a 2005 e 2008 a 2009. É importante ressaltar que o modelo conta com os fatores probabilísticos destes anos para a construção da matriz de transição. Como entrada o modelo solicita a quantidade inicial plantada por fase/índice de qualidade. Como resposta, após 32 ciclos, se tem a quantidade total da safra, indicativos de qualidade da cana colhida, morte e uma projeção da possível produção de açúcar e etanol com a quantidade colhida. Por fim, são apresentados os resultados das análises estatísticas onde os *outputs* do modelo são comparados com os existentes na base.

```
==============================
Modeling & Simulation - MCMC
Plantation: A2
Period: 2004 - 2005
========== INPUT ==========
B1 => 113 (t)
B2 => 43 (t)
B3 => 12 (t)
========== OUTPUT ==========
Col1 => 384(t)
Col2 => 421(t)
Col3 => 62 (t)
Mort => 23 (t)
Total => 867 (t)
---
Sugar + ethanol estimate => 112.71 (t) 12658.2 (L)
Ethanol estimate => 84272.4 (L)
========= STATISTIC =========
Correlation => 0.8872
p-value => 1.5534e-07
---
Shapiro
p-value => 0.6527
---
t => 0.013527
p-value => 0.8754
---
Kappa => 0.64
p-value => 0.00165
==============================
```

Figura 37 – Simulações MCMC para o período de 2004-2005

```
==============================
Modeling & Simulation - MCMC
Plantation: A2
Period: 2008 - 2009
========== INPUT ==========
B1 => 86 (t)
B2 => 67 (t)
B3 => 21 (t)
========== OUTPUT ==========
Col1 => 274(t)
Col2 => 579(t)
Col3 => 55 (t)
Mort => 32 (t)
Total => 908 (t)
---
Sugar + ethanol estimate => 118.04 (t) 13256.8 (L)
Ethanol estimate => 88257.6 (L)
========= STATISTIC =========
Correlation => 0.9167
p-value => 1.453e-06
---
Shapiro
p-value => 0.5383
---
t => 0.037854
p-value => 0.9129
---
Kappa => 0.68
p-value => 0.00143
==============================
```

Figura 38 – Simulações MCMC para o período de 2008-2009

É notado que mesmo já se tendo mapeado em banco os principais fatores que afetam no

ciclo da cana-de-açúcar, os resultados não são 100% precisos, porém foram satisfatórios. Ao se observar o coeficiente de correlação, o $p-value$ do teste de hipótese e o coeficiente Kappa, percebe-se que se teve uma representação aproximada do observado na realidade. A Figura 39 representa a simulação para o ano de 2015, em que os fatores ambientais não foram mapeados na base de simulação.

```
==============================
Modeling & Simulation - MCMC
Plantation: A2
Period: 2015 - 2016
=========== INPUT ===========
B1 => 95 (t)
B2 => 62 (t)
B3 => 8 (t)
=========== OUTPUT ==========
Col1 => 376(t)
Col2 => 441 (t)
Col3 => 47 (t)
Mort => 24 (t)
Total => 864 (t)
---
Sugar + ethanol estimate => 112.32 (t) 12614.4 (L)
Ethanol estimate => 83980.8 (L)
========= STATISTIC =========
Correlation => 0.8152
p-value => 1.836e-08
---
Shapiro
p-value => 0.4862
---
t => 0.56283
p-value => 0.7638
---
Kappa => 0.57
p-value => 0.00475
==============================
```

Figura 39 – Simulações MCMC para o período de 2015-2016

Mesmo com os fatores ambientais desconhecidos para esse período, pode-se perceber que os resultados ainda foram positivos. Foi observado um coeficiente Kappa de 0.57, com o modelo se baseando no histórico de anos anteriores. Se testando o modelo para um período mais a frente, de 2016 a 2017, também se obtém resultados satisfatórios. Como observado na Figura 40, o Kappa foi de 0.55, e o teste de hipótese e a análise correlacional também obtiveram bons resultados, como pode-se observar no $p-value$ do teste t e no *"Correlation"*.

Os resultados obtidos nas simulações demostram uma boa fidelidade do modelo quando aplicado a uma região mais complexa, até quando não se conhece os fatores daquele ano específico, deste de que se tenha um histórico da mesma região. Quando aplicado a uma região não mapeada sem fatores e histórico, considerando todos os resultados da base de 2004 a 2017 também se obteve um bom resultado, como pode ser visto na Figura 41, onde se tem um Kappa de 0.51, considerado como um resultado moderado. Porém, essa região não mapeada fica na mesma usina, logo os fatores são semelhantes. Para ambientes com condições diferentes é necessário se treinar o modelo antes.

```
==============================
Modeling & Simulation - MCMC
Plantation: A2
Period: 2016 - 2017
=========== INPUT ===========
B1 => 82 (t)
B2 => 74 (t)
B3 => 16 (t)
=========== OUTPUT ==========
Col1 => 308 (t)
Col2 => 529 (t)
Col3 => 58 (t)
Mort => 31 (t)
Total => 895 (t)
---
Sugar + ethanol estimate => 116.35 (t) 13067 (L)
Ethanol estimate => 86994 (L)
========= STATISTIC =========
Correlation => 0.7879
p-value => 1.437e-09
---
Shapiro
p-value => 0.6316
---
t => 0.07223
p-value => 0.7543
---
Kappa => 0.55
p-value => 0.00521
==============================
```

Figura 40 – Simulações MCMC para o período de 2016-2017

```
==============================
Modeling & Simulation - MCMC
Plantation: A5
Period: 2016 - 2017
=========== INPUT ===========
B1 => 30 (t)
B2 => 55 (t)
B3 => 10 (t)
=========== OUTPUT ==========
Col1 => 189 (t)
Col2 => 276 (t)
Col3 => 25 (t)
Mort => 12 (t)
Total => 490 (t)
---
Sugar + ethanol estimate => 63.7 (t) 7154 (L)
Ethanol estimate => 47628 (L)
========= STATISTIC =========
Correlation => 0.7162
p-value => 1.245e-07
---
Shapiro
p-value => 0.5268
---
t => 0.46237
p-value => 0.7133
---
Kappa => 0.51
p-value => 0.00775
==============================
```

Figura 41 – Simulações MCMC para uma região não mapeada

As Figuras 42, 43 e 44 ilustram a obtenção dos resultados da safra total, frente as 100

simulações realizadas para cada período, entre os anos de 2004 a 2005, 2008 a 2009 e 2015 a 2016. Se percebe que durante as iterações das simulações as probabilidades da matriz de transição são variadas, encontrando diferentes resultados. A média desses resultados é o que o modelo considera como resposta final da simulação.

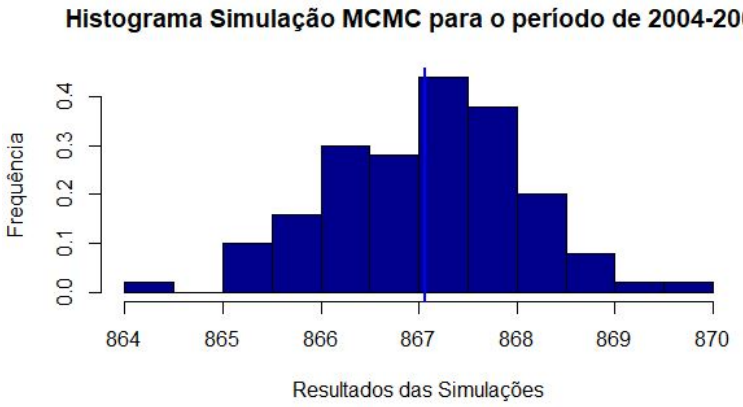


Figura 42 – Simulações realizadas para a determinação da safra no período de 2004-2005

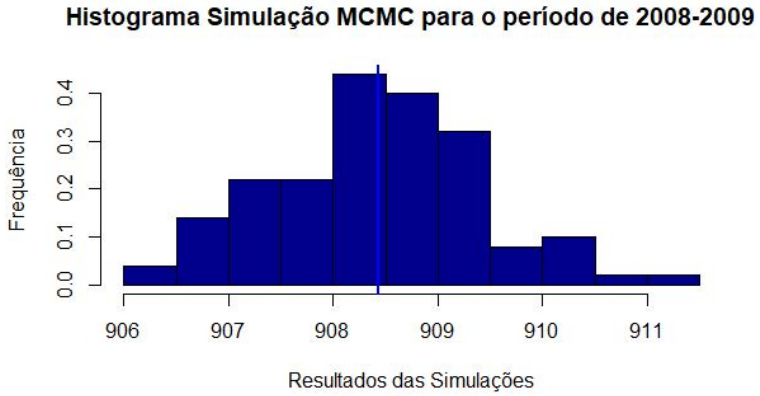


Figura 43 – Simulações realizadas para a determinação da safra no período de 2008-2009

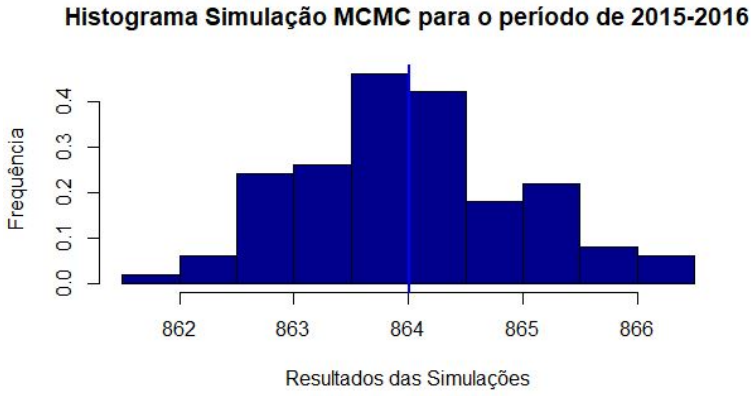


Figura 44 – Simulações realizadas para a determinação da safra no período de 2015-2016

A aplicação do modelo no contexto da área estudada na plantação trouxe resultados interessantes e animadores. O modelo conseguiu uma boa precisão de predição da quantidade total das safras, além de também conseguir fazer uma estimativa da qualidade das canas colhidas, entre boas, medianas e ruins. Essas informações são de grande interesse para a tomada de decisão por parte dos agrônomos e gestores da usina, podendo antecipar medidas para melhorias da safra ou a adaptação da indústria para a produção de distintos produtos originados da cana-de-açúcar.

O modelo para estimativa do total da safra foi mais assertivo que o de qualidade, porém, durante a revisão de literatura e nas entrevistas com os especialistas, não se identificou outros modelos computacionais para a predição das fases de crescimento da cana considerando seus índices de qualidade. Por se tratar de variáveis ordinais utilizou-se o Kappa ponderado para essas análises. A Tabela 11 ilustra os resultados da média dos índices Kappa obtidos se considerando os fatores de qualidade. Pode-se perceber que os resultantes ficam entre classificações Kappa de razoável a moderada.

Tabela 11 – Média kappa ponderado para índices de qualidade.

Período	**Kappa (Qualidade)**
2004 - 2005	0.45
2008 - 2009	0.47
2015 - 2016	0.42
2016 - 2017	0.38
Área não mapeada	0.32

Com essas validações, o modelo se apresenta como uma ferramenta promissora para a realização de simulações de cenários futuros dentro da usina mapeada. Esses resultados podem ser úteis para tomadas de decisão visando a otimização de processos e redução de custos. Porém, é importante se atentar, que mesmo o modelo conseguindo gerar resultados próximos a realidade, a grandeza utilizada é em toneladas, então uma pequena diferença numérica pode trazer impactos financeiros e estratégicos.

Outro aspecto que se pode notar é que se inserindo os dados em fases já avançadas do crescimento da cana-de-açúcar, como se tem menos ciclos temporais até a colheita, e pelos resultados estarem mais recentes e mais precisos, se obtém resultados finais com índice de confiança ainda melhor. Esse tipo de simulação também é bastante útil para a usina, pois pode simular cenários utilizando dados das fases atuais dos canaviais.

Apontar a previsão de um acontecimento não é uma ação trivial, sendo sujeito a diversas variações e erros. Visando a redução destes erros utilizou-se o MMC na realização das simulações das cadeias, o que ajudou a apontar resultados mais precisos. Por fim, a qualidade do modelo e sua precisão de acerto foram avaliadas pela obtenção do coeficiente de concordância de Kappa, validações cruzadas, por testes de hipóteses e pela opinião de especialistas da indústria sucroenergética, apresentando resultados positivos como um modelo piloto a ser implantado em usinas.

Para Searcy (1997) a ideia da agricultura de precisão é saber se o solo, dentre outras características relevantes para a produção, irão afetar a safra, assim podendo-se tomar medidas para a garantia da qualidade e também reduzindo custos, utilizando insumos só quando necessários. O modelo apresentado e validado neste capítulo visa a um futuro em que isso seja possível, que as probabilidades da matriz de transição possam estar sendo atualizadas em tempo real por sensores e pelos agrônomos e que o modelo consiga gerar simulações cada vez mais precisas e relevantes, relativas a safra e a sua qualidade.

7.3 Simulação e Experimentos com o Modelo de Autômatos Celulares Estocásticos

Como no modelo apresentado anteriormente, o modelo de ACE também foi idealizado para funcionar de forma genérica, podendo ser utilizado em diferentes plantações e regiões. O Capítulo 6 descreve o desenvolvimento deste modelo. Nesta seção será apresentado e discutido os resultados da utilização deste modelo em cenários e situações reais de análises na usina São José.

Em seu habitat natural, uma planta apresenta características relativas ao seu desenvolvimento e produção final, e quando é levada para um ambiente com condições climáticas diferentes, essas características podem ser modificadas. Portanto, tal fato mostra a necessidade de que cada região ou unidade realize estudos que possam avaliar o comportamento de variedades de cana-de-açúcar para uso em diferentes sistemas de produção (LUCCHESI, 1987). Para validar o modelo ACE nas condições da usina, inicialmente utilizou-se como representação os dados da área sob estudo de 15 hectares. Se seguiu o padrão de células de 25 m^2, e com isso se tem uma representação georreferenciada da área estudada através de um *grid* do autômato com 600 células, como ilustrado na Figura 45.

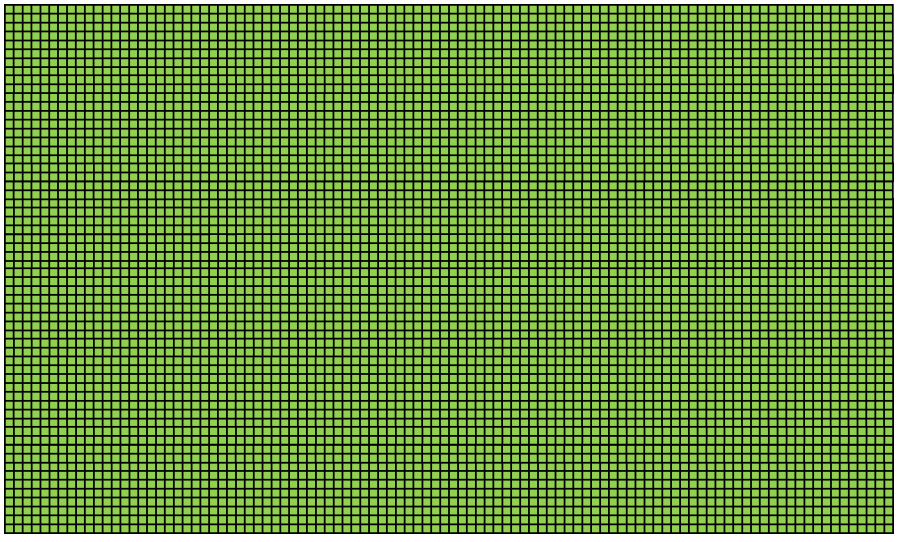


Figura 45 – Representação da área mapeada no AC

Como se pode perceber na Figura 45, devido a grande área analisada e ao tamanho reduzido das células, a visualização de toda a área do autômato fica comprometida. Por esse motivo, para ilustrar o resultado de alguns experimentos nesta seção serão montados modelos proporcionais de visualização mais simplificada.

A primeira lei da geografia afirma que tudo está relacionado a tudo mais, mas coisas próximas estão mais relacionadas do que coisas distantes (TOBLER, 1979), configurando-se assim o conceito de vizinhança. A dinâmica espaço-temporal das células e suas vizinhas é o que constitui as interações básicas nas simulações dos autômatos. Portanto, para as primeiras simulações realizadas, utilizou-se os dados do período entre 2004 e 2014 da área sob estudo. Os dados de 2015 a 2017 foram utilizados em um segundo momento como experimento de teste do modelo para uma área sem dados já pré-utilizados na modelagem das probabilidades de transição. Os experimentos foram realizados para todos os períodos existentes na base de dados, porém, só os mais representativos serão ilustrados e discutidos aqui.

Se aplicando as funções de transição de forma estocástica e em 32 ciclos temporais para o período de 2004 a 2005, se tem o cenário observado na Tabela 12 e na Figura 46. A Tabela 12 mostra a distribuição dos estados das células de acordo com a quantidade de ciclos temporais / fase de crescimento da cana-de-açúcar. A Figura 46 ilustra de forma simplificada esses cenários em AC.

Tabela 12 – Descrição de estados por ciclo temporal do autômato no período de 2004 a 2005.

Estado	**Brotação**	**Perfilhamento**	**Crescimento Vegetativo**	**Maturação**	**Colheita**
Boa	212	121	98	105	0
Mediana	327	385	392	403	0
Má	61	66	52	24	0
Morte	0	28	58	68	82
Colheita	0	0	0	0	518

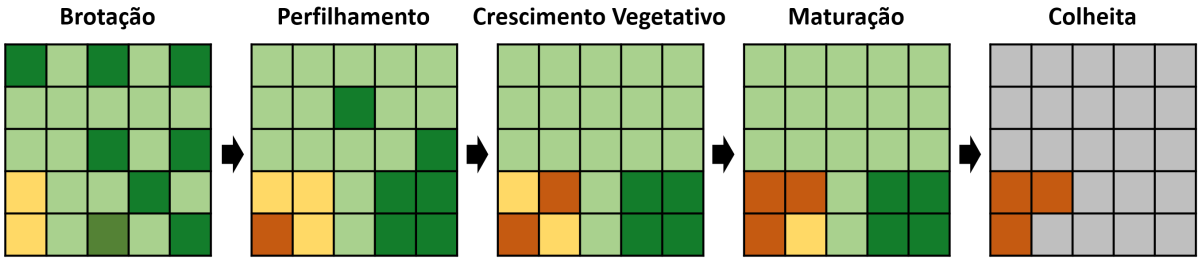


Figura 46 – Simplificação do AC no período de 2004 a 2005

Com os resultados obtidos do modelo se testou se há correlação com os existentes na base de dados, também se testou a normalidade das amostras, se aplicou o teste de hipóteses e por fim identificou-se o coeficiente kappa. Observou-se um coeficiente de correlação de 0.8465, com um $p - value = 1.931\text{e-}06$, indicando uma forte correlação. O

$p-value$ do teste Shapiro Wilk foi igual a 0.4654, indicando a normalidade da amostra. No teste de hipótese T Student se obteve um $p-value$ de 0.8635, indicando uma equivalência entre os resultados observados no modelo e nas bases. Por fim, o coeficiente Kappa foi de 0.71, considerando uma relação forte/substancial entre os resultados obtidos e os existentes na base de dados.

O período de 2004 a 2005 foi um dos utilizados para o treinamento do modelo. Logo, espera-se realmente que os resultados sejam mais precisos. Frente as métricas de correlação, teste de hipótese e índice Kappa aplicados, o modelo obteve um bom resultado na simulação deste cenário. Escolheu-se o período de 2015 a 2016 para se testar o modelo em relação a um período futuro, ainda não mapeado na base. A Tabela 13 e a Figura 47 ilustram esses resultados.

Tabela 13 – Descrição de estados por ciclo temporal do autômato no período de 2015 a 2016.

Estado	**Brotação**	**Perfilhamento**	**Crescimento Vegetativo**	**Maturação**	**Colheita**
Boa	289	241	212	187	0
Mediana	285	298	303	310	0
Má	26	37	33	42	0
Morte	0	24	52	61	73
Colheita	0	0	0	0	527

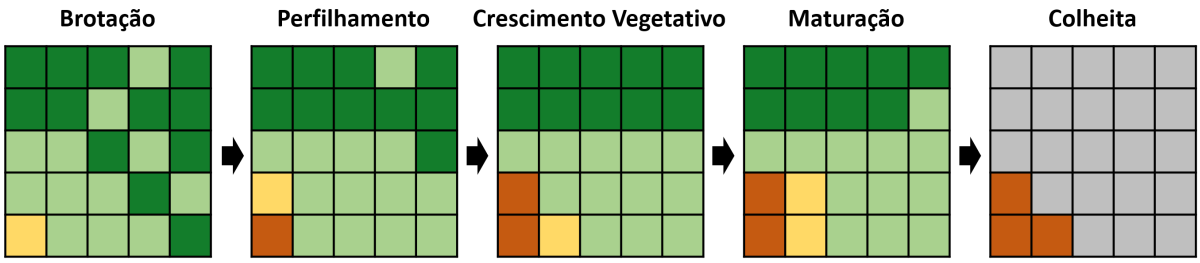


Figura 47 – Simplificação do AC no período de 2015 a 2016

Também se avaliou o coeficiente de correlação, teste de hipótese e Kappa para este experimento. O coeficiente de correlação foi de 0.6887, com um $p-value$ = 1.5535e-07, indicando uma correlação moderada. O $p-value$ do Shapiro Wilk foi igual a 0.5248, indicando a normalidade da amostra. No teste de hipótese T Student se obteve um $p-value$ de 0.7134, indicando uma equivalência entre as amostras. Por fim, o coeficiente Kappa foi de 0.62, o que ainda indica uma relação forte/substancial entre os resultados obtidos e os existentes na base de dados. O modelo não foi treinado com dados deste período, de 2015 a 2016, mesmo assim conseguiu um bom resultado frente as métricas analisadas.

As Figuras 46 e 47 representam uma simplificação do diagrama espaço-temporal de tamanho n=600 da área sob estudo do canavial. Porém, mesmo sendo simplificações, ao analisá-las percebe-se que regiões específicas do *grid*, mesmo em anos diferentes, apresentam o mesmo comportamento, improdutivo ou produtivo. Essa, por exemplo, é uma

informação georreferenciada relevante para gestores e agrônomos, pois pode-se traçar um plano para compreender o que está acontecendo nessas áreas e assim, propor possíveis soluções para mitigar as problemáticas.

Um conceito semelhante ao MMC utilizado no modelo apresentado na Seção 7.2 também foi utilizado neste modelo ACE. A utilização de iterações com fatores estocásticos no autômato, como no modelo anterior, possibilitaram uma sofisticação na obtenção dos resultados, o que levou a uma maior eficiência dos modelos. A introdução do método estocástico produz uma perturbação no sistema representada pela alteração do estado global, ela consegue retirar o autômato da inércia, o que retorna resultados mais próximos aos observados na realidade. É notável que em simulações determinísticas, em alguns passos de evolução, os autômatos ficam confinados em ciclos atratores, reproduzindo o comportamento estacionário da evolução determinística.

Outro teste realizado foi quanto a inserção de fatores externos nas simulações, como irrigação, adubo e pragas. Se realizou vários testes com esses fatores, se obtendo validações distintas. O coeficiente Kappa médio observado foi de 0.6, indicando uma relação moderada entre as simulações e o observado nas bases de dados. Com uma boa precisão na estimativa da utilização/existência desses fatores pode-se decidir sobre locais para se ter maior atenção de irrigação ou adubação, por exemplo, e locais mais estáveis, onde possivelmente as condições ambientais já atendem as necessidades, com isso pode-se otimizar recursos e reduzir custos.

Um comportamento interessante que foi observado é do que em regiões com pragas, quando cercadas por regiões saudáveis, com o estado "bom", mesmo estas podendo estar em áreas relativamente grandes, são controladas pelas regiões saudáveis que as cercam. As células saudáveis impedem a propagação de pragas nas simulações. Esta é uma hipótese observada nos resultados do modelo, esse fenômeno precisa ser validado em campo.

Quanto aos ciclos temporais, se observou que quanto maior a quantidade de ciclos, menos assertivo é o modelo, porém, para os 32 ciclos de uma safra de cana de ano e meio, se consegue uma boa precisão representativa. Depois de 48 ciclos, percebe-se que os resultados começam a ficar incoerentes. Assim como discutido anteriormente, é possível observar nos espaços temporais uma tendência do sistema voltar ao comportamento de um modelo determinístico no estado estacionário, onde a sequência de estados tende a cair em um ciclo atrator.

Também pode-se notar que estados semelhantes tendem a ficar agrupados ao invés de ficarem espalhados pelo *grid*. Isto pode ser interpretado como uma estratégia para ambientes comuns nas áreas geográficas, sendo mais fácil identificar setores a serem tratados. Conforme Antuniassi (1998), o mapeamento detalhado dos fatores de produção e aplicação localizada de insumos são os princípios básicos de um sistema agrícola eficiente.

Todos os fatores ambientais que impõem outra condição, tendem a alterar o ciclo fenológico da cultura. Portanto, todo o manejo deve ser empregado para otimizar e dar

condições para que a cana-de-açúcar consiga expressar o máximo do seu potencial produtivo, o que significa produzir quantidade de biomassa por hectare, aliado a quantidade de sacarose, preservando o canavial para ser produtivo o máximo de tempo possível, conseguindo-se assim colheitas mais lucrativas (SEGATO et al., 2006). A visualização, a qual permite um monitoramento das plantações, e as simulações permitem uma tomada de decisão com base em qualidade da plantação de forma georreferenciada, considerando diferentes fatores críticos para o plantio da cana-de-açúcar. As simulações que incluem fatores como irrigação e adubação são muito úteis para se prever e combater cenários de seca, assim otimizando as safras na hora das colheitas. Portanto, o modelo aqui apresentado obteve resultados satisfatórios para a análise e monitoramento de cenários de cultivo na cultura canavieira.

7.4 Avaliação e Discussão dos Resultados

Ambos os modelos apresentados nesta dissertação obtiveram resultados satisfatórios na modelagem e simulação de plantações de cana-de-açúcar. Foi possível realizar, com um bom índice de precisão, a predição de safras e seus índices de qualidade, bem como cenários de qualidade georreferenciados das plantações. Boas métricas se tratando da predição de um sistema complexo como é a dinâmica de um canavial.

Outro fator importante discutido neste trabalho é a apresentação de uma plataforma que une dispositivos IoT e computação em nuvem para a implantação de conceitos de agricultura de precisão e de indústria 4.0 nas usinas. Os modelos aqui apresentados também foram idealizados para poderem ser integrados a essa plataforma, possibilitando assim, um monitoramento em tempo real dos dados das plantações, bem como a realização de simulações com dados mais precisos e recentes.

Pode-se comparar os modelos desenvolvidos com outros encontrados na literatura e utilizados na usina. O modelo de Scarpari (2002) utiliza dados climáticos, como temperatura e precipitações, para a predição de TCH e Açúcar Teórico Recuperável (ATR). O modelo proposto consegue bons resultados na predição de ATR na Usina Bortolo Carolo, porém afirma que o modelo para predição de TCH não se mostraram muito confiáveis.

Pacheco (2005) utiliza um modelo de redes neurais artificiais para a predição de Pol da Cana Corrigido (PCC), TCH e fibra em canaviais. O modelo apresenta resultados satisfatórios na predição de TCH, com 78% de acerto e na predição de fibras na casa dos 90%. O trabalho não realiza a predição por qualidade da safra colhida, porém destaca a importância de um modelo capaz de o realizar. Trigo (2005) utiliza o trabalho de Pacheco como base e consegue melhorar a predição do TCH para 79,5%. A sua predição PCC fica abaixo da de Pacheco e o supera novamente na predição de fibras.

Como observado na etapa de imersão desta pesquisa, o modelo mais utilizado para predição de safras pelas usinas é o estatístico inferencial, no qual com base em dados

gerais do passado, se aplica alguma técnica mais trivial via Excel para a obtenção de uma estimativa de situações futuras. Alguns setores da usina ainda utilizam a predição embasada em conhecimentos e experiência de funcionários antigos, que consegue estimar cenários via observações e acompanhamento do canavial.

Os modelos apresentados e discutidos neste trabalho apresentaram bons resultados para a predição de TCH, TPH e safra localizada, além de se conseguir uma visualização georreferenciada de possíveis cenários de qualidade futuros da plantação. Porém, percebe-se que os demais modelos encontrados na literatura foram aplicados a todo o canavial. Neste estudo, apenas a simulação por hectare considerou toda a área do canavial, tendo uma boa estimativa. No entanto, os demais experimentos foram referentes a uma área específica da plantação, onde se tinha conhecimento sobre distintas variáveis que influenciam na qualidade e produção do canavial. Outro ponto a ser considerado é que os experimentos foram realizados considerando a unidade de medida em toneladas, logo, perdas que podem aparentar ser pequenas em escala numérica, são grandes dado a unidade de medida utilizada. Pode-se notar também que os modelos são muito sensíveis a quantidade, refino e ajustes dos dados, isso é refletido em uma dependência básica para o bom funcionamento destes.

Os resultados aqui apresentados são promissores, porém é interesse que o modelo seja validado em uma área ainda maior e em outras regiões e usinas. A contribuição parte de um estudo que envolveu uma metodologia para compreensão da problemática e identificação dos principais fatores que influenciam no ciclo e qualidade dos canaviais. Esta contribuição une-se a um modelo de MD, utilização de MCMC e de ACE para predições de cenários e safras de forma significante, com a grande diferenciação dos demais modelos identificados na literatura, de forma georreferenciada e considerando índices de qualidade.

Um outro conceito vem sendo testado neste trabalho, a união dos dois modelos propostos, em que se utiliza o ACE para a simulação de um cenário da plantação após alguns ciclos, já que esse modelo é mais preciso quanto a predição de qualidade e pode-se utilizar os *outputs* gerados por ele como *inputs* na cadeia de Markov para a determinação do total da safra. A realização desse experimento vem trazendo resultados interessantes, porém, por não estarem concluídos serão incluídos em trabalhos futuros para que na continuação da pesquisa se possa gerar um artigo científico com esse *ensemble*, se for comprovado como viável.

A Figura 48 ilustra uma representação geral do projeto. No (1) se têm os pontos para coleta de dados, oriundos das plantações, especialistas e de outras bases. No (2) os dados são armazenados e minerados para serem disponibilizados de forma estruturada e alimentar os modelos. No (3) os dados são utilizados como base para os modelos MCMC e ACE que retornam as informações e resultados das análises para em (4) poder se montar a visualização dos dados. Por fim, em 5 se tem a visualização das informações de interesse relativos ao total de safra, qualidade e total de produção estimada de açúcar e etanol.

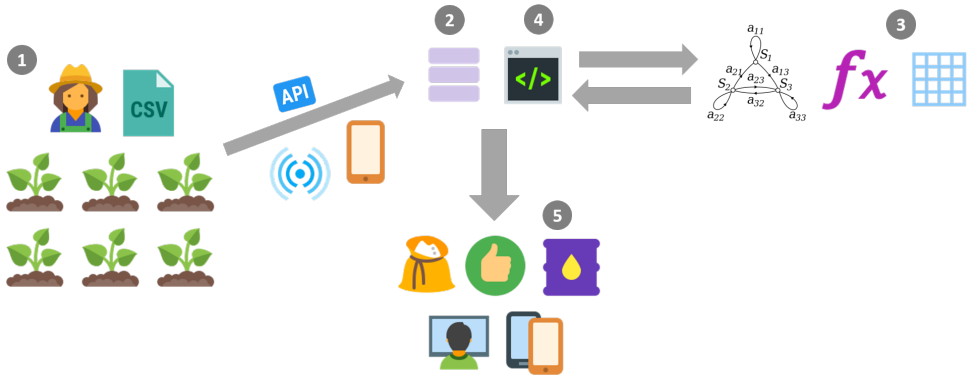


Figura 48 – Representação do projeto integrado

Capítulo 8

Conclusão

A agricultura de precisão é uma ferramenta de inovação tecnológica que, quando aplicada de forma preventiva durante a produção agrícola, melhora os aspectos econômicos e de gestão, minimizando perdas de insumos e riscos ambientais em relação à agricultura convencional (MACHADO et al., 2018). Organizações internacionais argumentam que o investimento em tecnologias agrícolas e uma transição para a agricultura de precisão é uma tarefa obrigatória para garantir o suprimento de alimentos para nove bilhões de pessoas até 2050 (TAYLOR, 2018).

Como pode-se observar, o investimento em pesquisas, inovações e tecnologias para o setor agrário é de grande importância para diferentes áreas sociais. Este trabalho apresenta para a comunidade propostas de modelos para a modelagem e simulação de plantações de cana-de-açúcar, focados na predição de safras e qualidades do cultivo. As informações extraídas dos modelos podem ser úteis para a otimização da produção e redução de impactos ambientais por parte da agroindústria sucroenergética.

Para a realização da pesquisa e desenvolvimento do projeto, foi preciso realizar uma imersão na realidade e problemáticas enfrentadas dentro de uma usina. Foi realizado um estudo de caso na São José Agroindustrial, uma das maiores usinas sucroenergéticas do estado de Pernambuco. No estudo de caso foi possível compreender as necessidades de agrônomos, gestores e pesquisadores da usina e elaborar soluções que atendessem aos requisitos solicitados.

Os resultados obtidos só foram possíveis graças aos processos metodológicos aplicados. A utilização do *Toolkit* HCD foi fundamental para a imersão, coleta de dados e compreensão da problemática. A solução foi desenvolvida em conjunto com os *stakeholders* e pode ser validada com dados da usina e por especialistas. O CRISP-DM foi de grande relevância para as fases de MD, pois com a aplicação desta metodologia o processo para a identificação de fatores chaves e correlacionados a cada fase de crescimento da cana-de-açúcar foi possível com maior facilidade. Com isso, foi possível determinar as principais variáveis que influenciavam na safra e qualidade do cultivo.

A plataforma em nuvem e infraestrutura propostas neste trabalho são sugestões para

o desenvolvimento de usinas mais modernas, conectadas e sustentáveis. A utilização da IoT nas indústrias já vem demostrando avanços importantes na onda de inovação trazida pela indústria 4.0 e a utilização dessas tecnologias também vem ganhando força no cenário das agroindústrias, visando ao monitoramento e a otimização de processos alinhados a preservação do meio ambiente. As propostas aqui apresentadas já visam a aplicação dessas novas tecnologias nas usinas sucroenergéticas.

O modelo MCMC demostrou resultados promissores nas simulações para a predição de safras, TCH, TPH e índices de qualidade por fase de crescimento da cana-de-açúcar. As técnicas de validação aplicadas ao experimento demostram um potencial para uma solução que pode agregar valores importantes a agroindústria. Através da predição de safras por qualidade pode se tomar, ou não, ações para otimizar a produção. Além disso, dado a safra estimada, pode-se decidir sobre o conjunto de produtos que serão gerados a partir do quantitativo e qualidade da cana-de-açúcar colhida.

O modelo ACE trouxe uma proposta de contribuição importante para a área de modelagem de canaviais, sendo um modelo para o mapeamento georreferenciado da plantação, considerando índices de qualidade das diferentes fases de crescimento da cana-de-açúcar. O ACE pode apresentar aos gestores e agrônomos um monitoramento em tempo real da plantação, além de possibilitar simulações por áreas específicas do canavial, podendo testar cenários com a aplicação de irrigação, ou surtos de alguma praga. A avaliação deste modelo, assim como o MCMC, se deu por uma validação cruzada com os dados observados nas bases utilizadas. O modelo ACE também apresentou resultados significativos nos experimentos realizados, analisando-se os valores encontrados pelas métricas abordadas.

Foi fundamental para o desenvolvimento da pesquisa a presença do público alvo em todas as etapas do projeto, garantindo que a solução final fosse útil e atendesse as suas principais necessidades. Além dos indicadores obtidos nos experimentos, o mais satisfatório foi os *feedbacks* positivos dos *stakeholders* em relação ao projeto. O estudo apresenta para a comunidade científica modelos para a predição de safras e cenários de qualidade do cultivo, tudo isso relacionado a uma proposta de plataforma em nuvem para armazenamento, coleta e compartilhamento dos dados e informações.

Em matéria recente do Diário de Pernambuco (2019) se tem a notícia de que a moagem da cana-de-açúcar na safra 2018/2019, que se encerra no fim do próximo mês, será a melhor safra dos últimos quatro anos para o estado. Essa melhora é atribuída ao investimento em tecnologias e inovações por parte das usinas e por condições climáticas favoráveis. Houve aumento na produção de cana, 11 milhões de toneladas (aumento de 8,83%), açúcar, 695 mil toneladas e etanol (aumento de 8%), 415 milhões de litros (aumento de 28%).

O agronegócio não é importante apenas para o Brasil, mas para o mundo. Sendo assim, processos, inovações, pesquisas e investimentos devem ser realizados e são necessários para garantir que este importante setor consiga crescer de forma sustentável, garantindo produtos saudáveis, preservação do meio ambiente, qualidade do trabalho e uma produção

otimizada que atenda a demanda global. Portanto, é preciso que governo e organizações privadas invistam em pesquisas e inovações, tanto para melhorarem os seus processos, como as comunidades as quais estão inseridos. É preciso garantir a consciência da importância do governo apoiar programas de pesquisa e financiamentos para estudantes, universidades e pequenos produtores rurais, assim garantindo avanços importantes para a agricultura nacional e para a sua sociedade.

8.1 Principais Contribuições

Frente aos resultados obtidos durante o desenvolvimento deste projeto, pode-se constatar as reais contribuições da realização desta pesquisa. Com isso, pode-se apresentar a comunidade:

- ❏ Um levantamento de trabalhos científicos, técnicas computacionais, modelos e ferramentas para modelagem e simulação de monoculturas. Bem como, uma visão macro do estado da arte em pesquisas envolvendo agricultura de precisão e computação;
- ❏ Uma sugestão de plataforma em nuvem para coleta, armazenamento, monitoramento e análise dos dados das plantações;
- ❏ Proposta de um modelo baseado em cadeias de Markov de Monte Carlo para a modelagem de plantações de cana-de-açúcar visando a previsão de safra e qualidade;
- ❏ Proposta de um modelo baseado em autômatos celulares estocásticos aplicados à modelagem dinâmica de plantações de cana-de-açúcar, permitindo a simulação de cenários georreferenciados;
- ❏ Proposta de uma plataforma que integra o uso de metodologias como o *Toolkit* HCD e o CRISP-DM com diferentes modelos e tecnologias matemático-computacionais, para a modelagem e simulação de safras e qualidade destas visando a otimização de processos da indústria sucroenergética.

Todas as sugestões e propostas resultantes desta pesquisa visam a fornecer informações estratégicas para a tomada de decisão por parte dos gestores das usinas. Essas informações podem ser utilizadas como instrumentos para a otimização das safras e como apoio para a redução de possíveis impactos ambientais.

8.2 Trabalhos Futuros

O projeto e os modelos propostos apresentados neste estudo apresentaram resultados satisfatórios frente as plantações e dados analisados. Os resultados foram promissores

como indicativos de uma ferramenta eficiente para o monitoramento e simulação de cenários nas plantações de cana-de-açúcar. Contudo, para uma plena validação dos modelos seria interessante a aplicação destes em outros canaviais, se possível de outras regiões do país, com diferentes condições ambientais. Uma validação em diferentes canaviais e com diferentes condições ambientais seria fundamental para decretar a validade dos modelos, como ferramentas genéricas e eficientes para serem implantadas em diferentes plantações para a modelagem e simulação dos processos de cultivo da cana-de-açúcar.

Os modelos também podem ser otimizados, validando mais variáveis e fatores que influenciam, ou não, nas fases de crescimento do canavial. Uma determinação precisa desses fatores chaves significa um determinante na melhora dos resultados gerados pelos modelos. Mais estudos e observações desses fatores podem trazer resultados mais precisos e assertivos.

Seria interessante melhorar os modelos para integrarem um cruzamento entre a qualidade da safra com a viabilidade da produção de etanol ou açúcar, servindo para a tomada de decisão de qual se produzir. Hoje os modelos estimam a produção de açúcar e etanol embasados no total da safra estimado, também seria interessante uma adaptação da análise do modelo que considerasse os impactos dos índices de qualidade estimados na produção do etanol e do açúcar.

O projeto foi desenvolvido com o estudo de caso aplicado na iniciativa privada, mas como trabalhos futuros apresenta características que mostram um grande potencial para ser disponibilizada como ferramentas *open source*. Pode-se perceber que a solução pode ser facilmente adequada para outros tipos de cultura. A disponibilização *open source* da plataforma daria a comunidade uma importante ferramenta para o compartilhamento de dados de diferentes plantações, possibilitando trocas mais intensas de conhecimentos e a possível criação de uma *crowd computing* de plantações.

Por fim, se tratando da modelagem de um sistema complexo e dinâmico, é preciso que haja um acompanhamento de todas as unidades que formam a rede complexa dos fatores que integram um canavial. Com isso, espera-se ter a contribuição de uma ferramenta ou de possíveis modelos que serão úteis para modelagens e simulações precisas dos diferentes cenários enfrentados em plantações de cana-de-açúcar.

8.3 Contribuições em Produção Bibliográfica

Durante o período de mestrado foram realizadas oito publicações entre eventos nacionais e internacionais, todos classificados e em diferentes áreas. Essas pesquisas foram resultado das disciplinas cursadas e pesquisas que serviram para embasar conceitos e métodos utilizados na pesquisa apresentada nesta dissertação. Os artigos apresentados durante este período foram esses:

- CRUZ JÚNIOR, G. G. ; NASCIMENTO, R. L. S. ; CYSNEIROS FILHO, G. A.

A. ; ROLIM, V. B. ; SANTOS, E. ; ALVES, G. . Planejamento de serious games focados nos usuários e a serviço da preservação dos biomas brasileiros. In: Simpósio Brasileiro de Jogos e Entretenimento Digital, 2017, Curitiba. SBC Proceedings of SBGames 2017 - Art & Design Track. Curitiba: SBC Proceedings of SBGames 2017, 2017. v. 1. p. 402-405.

- CRUZ JÚNIOR, G. G. ; NASCIMENTO, R. L. S. ; GOUVEIA, R. M. M. ; ALVES, G. . Um serious game multiplataforma para o ensino e difusao da cultura da reciclagem. In: Simpósio Brasileiro de Jogos e Entretenimento Digital, 2017, Curitiba. SBC Proceedings of SBGames 2017 - Culture Track. Curitiba: SBC Proceedings of SBGames 2017, 2017. v. 1. p. 922-929.

- CRUZ JÚNIOR, G. G. ; NASCIMENTO, RAFAELLA ; CYSNEIROS, GILBERTO ; ALVES, GABRIEL ; SANTOS, EDNILZA . Internet das Coisas, Games e Data Science a serviço da conscientização e preservação da Caatinga. In: XXVIII Simpósio Brasileiro de Informática na Educação SBIE (Brazilian Symposium on Computers in Education), 2017, Recife. org.crossref.xschema.1.Title@19a8e580, 2017. p. 714.

- CRUZ JÚNIOR, G. G. ; NASCIMENTO, RAFAELLA ; ALVES, GABRIEL ; GOUVEIA, ROBERTA . Identificando Correlações e Outliers Entre Bases de Dados Educacionais. In: VI Congresso Brasileiro de Informática na Educação, 2017, Recife. org.crossref.xschema.1.Title@554d0930, 2017. p. 694.

- NASCIMENTO, R. L. S. ; CRUZ JÚNIOR, G. G. ; CRUZ JÚNIOR, G. G. ; GOUVEIA, R. M. M. . Mineração De Dados Abertos Da Educação Básica E Análise Com Mapas De Calor. In: XV Congresso Internacional de Tecnologia na Educação, 2017, Recife. Anais do XV Congresso Internacional de Tecnologia na Educação, 2017.

- CRUZ JÚNIOR, G. G. ; NASCIMENTO, RAFAELLA ; CARNEIRO, NADJA ; LIMA, RINALDO . Design Thinking & Comunicação Aumentativa e Alternativa como ferramentas para o ensino e auxílio de professores do Atendimento Educacional Especializado. In: XXIX Simpósio Brasileiro de Informática na Educação (Brazilian Symposium on Computers in Education), 2018, Fortaleza. org.crossref.xschema.1. Title@13c76b08, 2018. p. 1173.

- NASCIMENTO, RAFAELLA ; CRUZ JÚNIOR, G. G. . Estudo sobre Docentes do Ensino Básico através de Indicadores Educacionais e Modelos de Regressão. In: VII Congresso Brasileiro de Informática na Educação, 2018, Fortaleza. org.crossref.xschema.1.Title@20a45d6f, 2018. p. 379.

- NASCIMENTO, R. L. S. ; CRUZ JÚNIOR, G. G. ; FAGUNDES, R. A. A. . Mineração de Dados Educacionais: Um Estudo Sobre Indicadores da Educação em Bases de

Dados do INEP. RENOTE. REVISTA NOVAS TECNOLOGIAS NA EDUCAÇÃO, v. 16, p. 1, 2018.

Quanto as publicações diretamente relacionadas a essa pesquisa, utilizou-se a estratégia de que cada contribuição aqui proposta, gerasse um artigo para a sua fundamentação. No caso foram planejadas cinco publicações. Três destas já estão concluídas e foram submetidas para periódicos internacionais, são elas:

- CRUZ JÚNIOR, G. G. ; ALVES, GABRIEL. A survey of computational techniques, models and tools for monoculture modeling and simulation.
- CRUZ JÚNIOR, G. G. ; ALVES, GABRIEL. Cloud Plantations: Sharing Data in Cellular Automata for Agroindustry Research.
- CRUZ JÚNIOR, G. G. ; ALVES, GABRIEL. Stochastic Cellular Automata Applied to Dynamic Modeling of sugarcane plantations.

Outras duas publicações estão em processo de escrita:

- CRUZ JÚNIOR, G. G. ; ALVES, GABRIEL. Modeling sugarcane plantations with Monte Carlo Markov Chains for crop and quality prediction
- CRUZ JÚNIOR, G. G. ; ALVES, GABRIEL. Modeling and Simulation of Plantations for Agroindustry Optimization of Sugarcane.

O projeto final e seus resultados foram convidados a serem apresentados na 27º Feira Internacional de Tecnologia Sucroenergética (FENASUCRO & AGROCANA) que acontecerá entre os dias 20 e 23 de agosto desse ano em Sertãozinho, São Paulo. A FENASUCRO & AGROCANA é o maior evento do setor sucroenergético no mundo.

Referências

ADAMI, S. F. Autômatos celulares e sistemas de informações geográficas aplicadas à modelagem da dinâmica espacial da cana-de-açucar na região de araçatuba-sp. Tese (doutorado) - Universidade Estadual de Campinas, Instituto de Geociências, Campinas, SP., 2011.

ALMEIDA, R. M. et al. Autômatos celulares probabilísticos aplicados à modelagem da propagação de incêndios de vegetação. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 3, n. 1, 2015.

ALVES, A.; SOUZA, F. d.; MARQUES, M. Avaliação do potencial à erosão dos solos: uma análise comparativa entre lógica fuzzy e o método usle. **Anais. XII Simpósio Brasileiro de Sensoriamento Remoto, Goiânia**, v. 1, p. 2011–2018, 2005.

ALVES, R.; DELGADO, C. Processos estocásticos. Faculdade de Economia, Universidade do Porto, 1997.

AMO, S. de. Técnicas de mineração de dados. **Jornada de Atualizaçao em Informatica**, 2004.

ANTUNIASSI, U. Agricultura de precisão: aplicação localizada de agrotóxicos. **Tecnologia e segurança na aplicação dos agrotóxicos-novas tecnologias. Santa Maria: Departamento de Defesa Fitossanitária**, p. 53–63, 1998.

ARALDI, A. A. R. Medidas de tendência central e separatrizes. **CAV-UDESC, Lages-SC**, 2005.

AUDE, M. I. da S. Estádios de desenvolvimento da cana-de-açúcar e suas relações com a produtividade. **Ciência rural**, Directory of Open Access Journals, v. 23, n. 2, p. 241–248, 1993.

AYRES, F. Teoría y problemas de cálculo diferencial e integral. Mc Graw-Hill, 1978.

BAGNOLI, F. Cellular automata. **Bagnoli F, Ruffo S (eds) Dynamical Modeling in Biotechnologies. World Scientific, Singapore, p. 1**, 1998.

BALDUZZI, D.; TONONI, G. Integrated information in discrete dynamical systems: motivation and theoretical framework. **PLoS computational biology**, Public Library of Science, v. 4, n. 6, p. e1000091, 2008.

BANK, A. D. et al. The future of work: regional perspectives. 2018.

BAR-YAM, Y. Unifying principles in complex systems. **Converging Technology (NBIC) for Improving Human Performance, MC Roco and WS Bainbridge, Dds., Kluwer**, 2003.

BASILI, V. R.; CALDIERA, G.; ROMBACH, D. H. **The Goal Question Metric Approach. Encyclopedia of Software Engineering 1 (1994)**. [S.l.]: Wiley, 1994.

BASTOS, J. L. D.; DUQUIA, R. P. Medidas de dispersão: os valores estão próximos entre si ou variam muito. **Scientia Medica**, v. 17, n. 1, p. 40–44, 2007.

BATTY, M.; COUCLELIS, H.; EICHEN, M. **Urban systems as cellular automata**. [S.l.]: SAGE Publications Sage UK: London, England, 1997.

BAUER, L. Estimação do coeficiente de correlação de spearman ponderado. Dissertação - Programa de Pós-Graduação em Epidemiologia. Universidade Federal do Rio Grande do Sul, Porto Alegre - RS., 2007.

BERTSEKAS, D. P.; TSITSIKLIS, J. N. **Introduction to probability**. [S.l.]: Athena Scientific Belmont, MA, 2002. v. 1.

BIBI, S. et al. Ontology based bayesian software process improvenent. In: IEEE. **Software Engineering and Applications (ICSOFT-EA), 2014 9th International Conference on**. [S.l.], 2014. p. 568–575.

BOLDRINI, J. L. et al. **Álgebra linear**. [S.l.]: Harper & Row, 1980.

BORLIDO, D. J. A. Indústria 4.0: Aplicação a sistemas de manutenção. Dissertação - Mestrado Integrado em Engenharia Mecânica. Universidade do Porto, Faculdade de Engenharia., 2017.

BRÉMAUD, P. **Markov chains: Gibbs fields, Monte Carlo simulation, and queues**. [S.l.]: Springer Science & Business Media, 2013. v. 31.

BROWN, T. Definitions of design thinking. **Design Thinking: Thoughts by Tim Brown**, v. 7, 2008.

BRUINSMA, J. **World agriculture: towards 2015/2030: an FAO study**. [S.l.]: Routledge, 2017.

CÂMARA, G. Ecofisiologia da cultura da cana-de-açúcar. **CÂMARA, GMS; OLIVEIRA, EAM Produção de cana-de-açúcar. Piracicaba: FEALQ**, p. 31–64, 1993.

CANZIAN, E. et al. Projeto de um monitor de semeadora com gps para pesquisa em agricultura de precisão. **Disponível na Internet. http://www. pcs. usp. br/˜ laa/projetos. html em**, v. 27, 1999.

CASAGRANDE, A. A. **Tópicos de morfologia e fisiologia da cana-de-açúcar**. [S.l.]: Funep Jaboticabal, 1991. v. 3.

CASTRO, A.; LIMA, D. Autômatos celulares aplicados a modelagem de dinâmica populacional em situação de risco. In: **Workshop of Applied Computing for the Management of the Environment and Natural Resources**. [S.l.: s.n.], 2013.

CASTRO, M. L. A.; CASTRO, R. de O. Autômatos celulares: implementações de von neumann, conway e wolfram. **Revista de Ciências Exatas e Tecnologia**, v. 3, n. 3, p. 89–106, 2015.

CENTEC, I. C. d. E. T. Cadernos tecnológicos: Produtor de cana-de-açúcar. Edições Democrito Rocha, 2004.

CHANG, K.-C.; PEARSON, K.; ZHANG, T. Perron-frobenius theorem for nonnegative tensors. **Communications in Mathematical Sciences**, International Press of Boston, v. 6, n. 2, p. 507–520, 2008.

CHAPMAN, P. et al. Crisp-dm 1.0 step-by-step data mining guide. 2000.

CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. **The american statistician**, Taylor & Francis Group, v. 49, n. 4, p. 327–335, 1995.

COELHO, J. C. et al. Agricultura de precisão. **Prefácio, Lisboa**, 2004.

COELHO, P. M. N. Rumo à indústria 4.0. Dissertação - Mestrado em Engenharia e Gestão Industrial. Faculdade de Ciências e Tecnologia, Universidade de Coimbra, Portugal, 2016.

COHEN, J. **Statistical Power Analysis for the Behavioral Sciences. 2nd edn. Hillsdale, New Jersey: L**. [S.l.]: Erlbaum, 1988.

COSTA, C. C. d.; GUILHOTO, J. J. M.; IMORI, D. Importância dos setores agroindustriais na geração de renda e emprego para a economia brasileira. **Revista de Economia e Sociologia Rural**, SciELO Brasil, v. 51, n. 4, p. 787–814, 2013.

COSTA, C. da. Indústria 4.0: o futuro da indústria nacional. **POSGERE-Pós-Graduação em Revista/IFSP-Campus São Paulo**, v. 1, n. 4, p. 5–14, 2017.

DAHLIA, L. et al. Consumer preference for indigenous vegetables. **World Agroforestry Center**, 2009.

DANCEY, C. P.; REIDY, J. **Estatística sem matemática para psicologia**. [S.l.]: Penso Editora, 2013.

DIAKOSAVVAS, D. **Policy Instruments to Support Green Growth in Agriculture–Main Report**. [S.l.]: OECD, Trade and Agriculture Directorate, 2012.

ERMENTROUT, G. B.; EDELSTEIN-KESHET, L. et al. Cellular automata approaches to biological modeling. **Journal of theoretical Biology**, Elsevier Science, v. 160, n. 1, p. 97–133, 1993.

FAO. Commodity balances - crops primary equivalent. **Food and Agriculture Organization of the United Nations**, 2018.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

FERNÁNDEZ, R.; LOUIS, P.-Y.; NARDI, F. R. Overview: Pca models and issues. In: **Probabilistic Cellular Automata**. [S.l.]: Springer, 2018. p. 1–30.

FERREIRA, M. R. Autômato celular com probabilidades de transição dependentes da altura para o estudo do crescimento de superfícies. Dissertação - Pós-Graduação em Modelagem Matemática e Computacional. CEFET, Belo Horizonte - MG, 2009.

FIGUEIRA, M. M. C. Identificação de outliers. **Millenium**, Instituto Politéc de Viseu, 1998.

FIGUEIREDO-FILHO, D. B.; SILVA-JUNIOR, J. A. Desvendando os mistérios do coeficiente de correlação de pearson (r). **Revista Política Hoje**, v. 18, n. 1, 2010.

FILHO, E. B.; D'OTTAVIANO, I. M. L. Conceitos básicos de sistêmica. **Auto-Organização. Coleção CLE**, n. 30, p. 283–306, 2000.

FILHO, J. E. R. V.; FISHLOW, A. Agricultura e indústria no brasil: inovação e competitividade. Instituto de Pesquisa Econômica Aplicada (Ipea), 2017.

FILHO, T. K.; JUNIOR, J. Z.; LIMA, P. R. S. de R. Análise da transição entre dias secos e chuvosos por meio da cadeia de markov de terceira ordem. **Pesquisa Agropecuária Brasileira**, SciELO Brasil, v. 41, n. 9, p. 1341–1349, 2006.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: An overview. **AI magazine**, v. 13, n. 3, p. 57, 1992.

FU, S. C. Modelling epidemic spread using cellular automata. **University of Western Australia**, 2002.

GABRIEL, S. Cana tem melhor safra dos últimos quatro anos. In: . [S.l.]: Cana tem melhor safra dos últimos quatro anos. Diário de Pernambuco, Recife, 11 jan. 2019. Economia, p. A6., 2019.

GIDEY, E. et al. Cellular automata and markov chain (ca_markov) model-based predictions of future land use and land cover scenarios (2015–2033) in raya, northern ethiopia. **Modeling Earth Systems and Environment**, Springer, v. 3, n. 4, p. 1245–1262, 2017.

GLASSOP, D.; RAE, A. L.; BONNETT, G. D. Sugarcane flowering genes and pathways in relation to vegetative regression. **Sugar Tech**, Springer, v. 16, n. 3, p. 235–240, 2014.

GONÇALVES, A.; MARCONDES, M.; LAKATOS, E. Os testes de hipóteses como instrumental de validação da interpretação (estatística inferencial). **Marcondes MA, Lakatos EM. Técnicas em pesquisas. São Paulo: Atlas**, 1982.

GUILHOTO, J. et al. Pib da agricultura familiar: Brasil-estados. Ministério do Desenvolvimento Agrário (MDA). NEAD Estudos 19, 2011.

GUIMARÃES, P. R. B. Métodos quantitativos estatísticos. **Curitiba: IESDE Brasil SA**, 2008.

HAIR, J. F. et al. **Análise multivariada de dados**. [S.l.]: Bookman Editora, 2009.

HEYLIGHEN, F. Building a science of complexity. In: **1988 Annual Conference of the Cybernetic Society. Londin**. [S.l.: s.n.], 1988.

HOLDEN, N. M. et al. Review of the sustainability of food systems and transition using the internet of food. **npj Science of Food**, Nature Publishing Group, v. 2, n. 1, p. 18, 2018.

HROMKOVIČ, J. **Algorithmics for hard problems: introduction to combinatorial optimization, randomization, approximation, and heuristics**. [S.l.]: Springer Science & Business Media, 2013.

HUANG, J. et al. An integrated approach based on markov chain and cellular automata to simulation of urban land use changes. **Applied Mathematics & Information Sciences**, Natural Sciences Publishing Corp, v. 9, n. 2, p. 769, 2015.

HUANG, X.; LIN, J.; DEMNER-FUSHMAN, D. Evaluation of pico as a knowledge representation for clinical questions. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. **AMIA annual symposium proceedings**. [S.l.], 2006. v. 2006, p. 359.

IBGE. Indicadores agropecuários. pesquisas agropecuárias. **ibge.gov.br**, 2018.

IDEO. **Human Centered Design Toolkit**. [S.l.]: San Francisco, California/USA, 2014.

JESUS, L. F. de; SILVA, V. B. da; ROCHA, F. da G. Uso de software para detecção de doenças na cultura da soja com o auxílio de um drone autônomo. **Anais do Computer on the Beach**, p. 552–553, 2015.

JOSÉ, U. S. Infraestrutura - são josé agroindustrial. In: . [S.l.]: São José Agroindustrial. Infraestrutura. Disponível em: <http://saojoseagroindustrial.com.br/usina/infraestrutura>. Acesso em: 20 de mar. de 2018., 2018.

JR, N. F. K.; MCQUEEN, R. J.; BAKER, M. Learning and process improvement in knowledge organizations: a critical analysis of four contemporary myths. **The Learning Organization**, MCB UP Ltd, v. 3, n. 1, p. 31–41, 1996.

JUDGE, G. G.; SWANSON, E. R. Markov chains: Basic concepts and suggested uses in agricultural economics. **Australian Journal of Agricultural Economics**, Wiley Online Library, v. 6, n. 2, p. 49–61, 1962.

JUNIOR, P. S. P. et al. Uso de geotecnologias para avaliação multitemporal da cobertura vegetal e gerenciamento de uma região estuarina no nordeste do brasil. **UD y la geomática**, n. 10, p. 13–17, 2015.

JUVANHOL, R. S. **Modelagem da vulnerabilidade à ocorrência e propagação de incêndios florestais**. Dissertação (Mestrado) — Universidade Federal do Espírito Santo, 2014.

KAMPFF, A. J. C. Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. Tese (Doutorado) - Pós-Graduação em Informática na Educação. Centro de Estudos Interdisciplinares em Novas Tecnologias da Educação, Universidade Federal do Rio Grande do Sul, Porto Alegre - RS, 2009.

KAPOOR, A. et al. Implementation of iot (internet of things) and image processing in smart agriculture. In: IEEE. **Computation System and Information Technology for Sustainable Solutions (CSITSS), International Conference on**. [S.l.], 2016. p. 21–26.

KIMATI, H. et al. **Manual de fitopatologia: doenças das plantas cultivadas**. [S.l.]: Agronômica Ceres São Paulo, 1997. v. 2.

KITCHENHAM, B. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, n. 2004, p. 1–26, 2004.

KNAPP, S.; HEIJDEN, M. G. van der. A global meta-analysis of yield stability in organic and conservation agriculture. **Nature communications**, Nature Publishing Group, v. 9, n. 1, p. 3632, 2018.

LANZER, A. T. S. Um modelo de simulação de autômatos celulares para avaliação de condições de biodiversidade e resiliência na exploração de florestas naturais. Tese (doutorado) - Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina, Centro Tecnológico, Florianópolis - SC, 2004.

LAUDON KENNETH, L. J. Sistemas de informações gerenciais: Fundamentos da inteligência de negócios:gestão da informação e de banco de dados. **9º ed. São Paulo: ABDR.**, 2011.

LEITE, J. L. I.; CERQUEIRA, M. Autômatos celulares. Disponível em <http://www.cin.ufpe.br/ if114/Monografias/Automatos%20Celulares/especificacoes.htm>. Acesso em 05 jan. 2019., 2014.

LEITE, L. S. Autômatos celulares para otimizaçao de cenários em gerenciamento de recursos de energia. Dissertação (Mestrado) - Pós-Graduação em Informática Aplicada. Universidade Federal Rural de Pernambuco, Recife - PE, 2016.

LIMA, A. R. et al. Impactos da monocultura de eucalipto sobre a estrutura agrária nas regiões norte e central do espírito santo/impacts of eucalyptus monoculture on the agrarian structure in the northern and central regions of espírito santo. **REVISTA NERA**, n. 34, p. 12–36, 2017.

LIN, J. et al. Blockchain and iot based food traceability for smart agriculture. In: ACM. **Proceedings of the 3rd International Conference on Crowd Science and Engineering**. [S.l.], 2018. p. 3.

LIU, Y. **Modelling urban development with geographical information systems and cellular automata**. [S.l.]: CRC Press, 2008.

LOPES, M. A.; CONTINI, E. Agricultura, sustentabilidade e tecnologia. **Agroanalysis**, v. 32, n. 02, p. 27–34, 2012.

LÓPEZ, E. et al. Predicting land-cover and land-use change in the urban fringe: a case in morelia city, mexico. **Landscape and urban planning**, Elsevier, v. 55, n. 4, p. 271–285, 2001.

LOZANO, K. K. C. **Autômatos Celulares Probabilísticos com Aplicações a Sistemas Biológicos**. Tese (Doutorado) — Programa de Engenharia de Sistemas e Computação. Universidade Federal do Rio de Janeiro, Rio de Janeiro - RJ, 2017.

LUCCHESI, A. Fatores da produção vegetal. **Ecofisiologia da produção agrícola**, Potafos. Piracicaba, p. 1–9, 1987.

MACEDO, R. C.; ALMEIDA, C. M.; SANTOS, J. R. Modelagem dinâmica espacial da expansão da agricultura em campos novos-sc. **Geosul**, v. 33, n. 68, p. 260–285, 2018.

MACHADO, J. et al. Agricultura de precisão: Programas tecnológicos no brasil. **Revista Geama**, v. 4, n. 2, p. 23–30, 2018.

MANHÃES, C. M. C. et al. **Fatores que afetam a brotação e o perfilhamento da cana-de-açúcar**. [S.l.]: Vértices, 2015.

MANNEVILLE, P. et al. **Cellular Automata and Modeling of Complex Physical Systems: Proceedings of the Winter School, Les Houches, France, February 21–28, 1989**. [S.l.]: Springer Science & Business Media, 2012. v. 46.

MARCOS, W. P. Cadeia de markov aplicada ao manejo de pragas em lavoura cafeeira. Dissertação - Programa de Pós-graduação em Engenharia Mecânica. Universidade Federal de Uberlândia, Uberlândia - MG, 2015.

MARKO, K.; ZULKARNAIN, F.; KUSRATMOKO, E. Coupling of markov chains and cellular automata spatial models to predict land cover changes (case study: upper ci leungsi catchment area). In: IOP PUBLISHING. **IOP Conference Series: Earth and Environmental Science**. [S.l.], 2016. v. 47, n. 1, p. 012032.

MARÔCO, J. **Análise estatística com o SPSS Statistics**. [S.l.]: ReportNumber, Lda, 2011.

MARTIN, G. L.; EK, A. R. A comparison of competition measures and growth models for predicting plantation red pine diameter and height growth. **Forest Science**, Oxford University Press, v. 30, n. 3, p. 731–743, 1984.

MATA, J.; COHN, M. Cellular automata-based modeling program: synthetic immune system. **Immunological reviews**, Wiley Online Library, v. 216, n. 1, p. 198–212, 2007.

MATOS, P. F.; PESSOA, V. L. S. A modernização da agricultura no brasil e os novos usos do território. **Geo Uerj**, v. 2, n. 22, p. 290–322, 2011.

MCGILL, R.; TUKEY, J. W.; LARSEN, W. A. Variations of box plots. **The American Statistician**, Taylor & Francis Group, v. 32, n. 1, p. 12–16, 1978.

MEHTA, A.; PATEL, S. Iot based smart agriculture research opportunities and challenges. **Int. J. Technol. Res. Eng**, v. 4, p. 541–543, 2016.

MEKALA, M. S.; VISWANATHAN, P. A survey: Smart agriculture iot with cloud computing. In: IEEE. **Microelectronic Devices, Circuits and Systems (ICMDCS), 2017 International conference on**. [S.l.], 2017. p. 1–7.

MELOTTI, G. Aplicação de autômatos celulares em sistemas complexos: Um estudo de caso em espalhamento de epidemias. **MACSIN-UFMG, Belo Horizonte**, 2009.

MEYN, S. P.; TWEEDIE, R. L. **Markov chains and stochastic stability**. [S.l.]: Springer Science & Business Media, 2012.

MIRANDA, B. A. et al. Autômatos celulares aplicados à epidemiologia da esquistossomose em pernambuco-uma análise comparativa do processo de coleta de moluscos. In: **Anais do XXXI Congresso Nacional de Matemática Aplicada e Computacional. Belém: Sociedade Brasileira de Matemática Aplicada e Computacional**. [S.l.: s.n.], 2008. p. 630–6.

MIRANDA, C. d. F. Modelação linear de séries temporais na presença de outliers. Dissertação - Departamento de Matemática Aplicada. Faculdade de Ciências da Universidade do Porto, 2001.

MOORE, D. S. **The basic practice of statistics**. [S.l.]: WH Freeman New York, 2007. v. 2.

MORAIS, C. Escalas de medida, estatística descritiva e inferência estatística. Instituto Politécnico de Bragança, Escola Superior de Educação, 2005.

MORO, S.; LAUREANO, R.; CORTEZ, P. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In: EUROSIS-ETI. **Proceedings of European Simulation and Modelling Conference-ESM'2011**. [S.l.], 2011. p. 117–121.

MUNOZ-GARCIA J., M.-R. J. P.-A. A. Outliers: a formal approach. **International Statistical Review**, 1990.

NASCIMENTO, M. **O uso de simulação de Monte Carlo via Cadeias de Markov no melhoramento genético**. Tese (Doutorado) — Programa de Pós-Graduação em Estatística Aplicada e Biometria. Universidade Federal de Viçosa, Viçosa-MG, 2009.

NEAL, R. M. et al. Mcmc using hamiltonian dynamics. **Handbook of Markov Chain Monte Carlo**, v. 2, n. 11, p. 2, 2011.

NETO, D. L. d. A.; COSTA, E. d. F. et al. Dimensionamento do pib do agronegócio em pernambuco. **Brazilian Journal of Rural Economy and Sociology (Revista de Economia e Sociologia Rural-RESR)**, Sociedade Brasileira de Economia e Sociologia Rural, v. 43, n. 4, p. 1–33, 2005.

NORMANDO, A. D. C.; TJÄDERHANE, L.; QUINTÃO, C. C. A. A escolha do teste estatístico–um tutorial em forma de apresentação em powerpoint. Dental Press Journal of Orthodontics, v. 15, n. 1, p. 101–106, 2010.

O IDRISI. **Rio Grande do Sul, Universidade Federal do Rio Grande do Sul. . Disponível em: <https://www.ufrgs.br/labgeo/index.php/cr-idrisi/23-o-idrisi>. Acesso em: 10 maio. 2018.**, 2016.

OLIVEIRA, E. C. de. Comparação das diferentes técnicas para a exclusão de "outliers". ENQUALAB – 2008. Congresso da Qualidade em Metrologia, 2008.

OLIVEIRA, G. S.; VIANNA, G. K. Sistomate: Sistema inteligente de suporte à decisão no auxílio ao combate da requeima em culturas de tomate. 2015.

PACHECO, D. F.; REGUEIRA, F. S.; NETO, F. B. d. L. Utilização de redes neurais artificiais em colheitas de cana-de-açúcar para predição de pcc, tch e fibra. **Revista Alcoolbrás, São Paulo**, 2005.

PARK, S.; WAGNER, D. F. Incorporating cellular automata simulators as analytical engines in gis. **Transactions in GIS**, Wiley Online Library, v. 2, n. 3, p. 213–231, 1997.

PEIXOTO, M.; BARROS, L.; BASSANEZI, R. Um modelo de autômatos celulares para o espalhamento geográfico da morte súbita dos citros com parâmetro fuzzy. Biomatemática XIII, p. 67–73, 2003.

PONTES, A. C. F. Ensino da correlação de postos no ensino médio. **SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA (SINAPE)**, v. 19, p. 26–30, 2010.

PREECE, J.; ROGERS, Y.; SHARP, H. **Interaction design: beyond human-computer interaction**. [S.l.]: John Wiley & Sons, 2015.

RAO, M. P. V. et al. Smart agriculture monitoring system based on internet of things. 2018.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações**. [S.l.]: Editora Manole Ltda, 2003.

REZENDE, S. O. et al. Mineração de dados. **Sistemas inteligentes: fundamentos e aplicações**, v. 1, p. 307–335, 2003.

ROBERT, C.; CASELLA, G. **Monte Carlo statistical methods**. [S.l.]: Springer Science & Business Media, 2013.

RODRIGUES, R. B. R. L. **USO DE ADITIVOS NA QUALIDADE BROMATÓLOGICA DA SILAGEM DE CANA-DE-AÇÚCAR**. [S.l.]: FAFRAM, 2017.

ROSS, S. M. Stochastic processes. john wiley& sons. **New York**, 1983.

ROUSSEAU, G. et al. Dynamical phases in a cellular automaton model for epidemic propagation. **Physica D: Nonlinear Phenomena**, Elsevier, v. 103, n. 1-4, p. 554–563, 1997.

ROY, S. et al. Iot, big data science & analytics, cloud computing and mobile app based hybrid system for smart agriculture. In: IEEE. **Industrial Automation and Electromechanical Engineering Conference (IEMECON), 2017 8th Annual**. [S.l.], 2017. p. 303–304.

ROZA, D. Novidade no campo: Geotecnologias renovam a agricultura. **Revista InfoGEO, n**, 2000.

SANTAELLA, L. et al. Desvelando a internet das coisas. **Revista GEMInIS**, v. 4, n. 2, p. 19–32, 2013.

SCARPARI, M. S. **Modelos para a previsão da produtividade da cana-de-açúcar (Saccharum spp.) através de parâmetros climáticos.** Tese (Doutorado) — Universidade de São Paulo, Piracicaba - SP, 2002.

SEARCY, S. W. **Precision farming: A new approach to crop management**. [S.l.]: Texas Agricultural Extension Service, Texas A & M University System, 1997.

SEGATO, S. V. et al. Atualização em produção de cana-de-açúcar. **Piracicaba: CP**, v. 2, p. 415, 2006.

SEMBOLONI, F. An urban and regional model based on cellular automata. **Environment and Planning B: Planning and Design**, SAGE Publications Sage UK: London, England, v. 24, n. 4, p. 589–612, 1997.

SHEATS, R. D.; PANKRATZ, V. S. Understanding distributions and data types. In: ELSEVIER. **Seminars in Orthodontics**. [S.l.], 2002. v. 8, n. 2, p. 62–66.

SIEGEL, S.; CASTELLAN-JR, N. J. **Estatística não-paramétrica para as ciências do comportamento**. [S.l.]: McGraw-Hill São Paulo, 1975.

SILVA, M. H. M. da; RAKOCEVIC, M. Programa computacional para interpolação de crescimento vegetativo de plantas da erva-mate em 3d. **Pesquisa Agropecuária Brasileira**, v. 45, n. 3, p. 244–251, 2011.

SILVA, S. C. da; FERREIRA, R. A. Aspectos jurídicos e ambientais da monocultura da cana de açúcar. **Dat@ venia**, v. 9, n. 1, p. 112–124, 2018.

SILVEIRA-JÚNIOR, A. A. Aplicação das cadeias de markov no estudo do controle biológico da planta aquática eichhornia azurea. Dissertação - Mestrado profissional em Matemática. Universidade Federal de Goiás, Jataí, 2014.

SIMON, C. P.; BLUME, L.; DOERING, C. I. **Matemática para economistas**. [S.l.]: Bookman, 2004.

SOUZA, G. M.; BUCKERIDGE, M. S. Sistemas complexos: novas formas de ver a botânica. **Revista Brasileira de Botânica**, SciELO Brasil, v. 27, n. 3, p. 407–419, 2004.

SPEARMAN, C. The proof and measurement of association between two things. **The American journal of psychology**, JSTOR, v. 15, n. 1, p. 72–101, 1904.

SUMA, D. N. et al. Iot based smart agriculture monitoring system. **International Journal on Recent and Innovation Trends in Computing and Communication**, v. 5, n. 2, p. 177–181, 2017.

TAYLOR, M. Climate-smart agriculture: what is it good for? **The Journal of Peasant Studies**, Taylor & Francis, v. 45, n. 1, p. 89–107, 2018.

TEIXEIRA, L. M. et al. Projeção da dinâmica da floresta natural de terra-firme, região de manaus-am, com o uso da cadeia de transição probabilística de markov. **Acta amazonica**, Instituto Nacional de Pesquisas da Amazônia, 2007.

TEOFILO, D. **KDD – Knowlegde Discovery in Database**. 2015. <https://danielteofilo.wordpress.com/2015/02/16/kdd-knowlegde-discovery-in-database>. Acessado: 16-05-2016.

THOMAS, A. L. Desenvolvimento das plantas de batata, mandioca, fumo e cana-de-açúcar. UFRGS, Porto Alegre - RS, 2016.

TOBLER, W. R. Cellular geography. In: **Philosophy in geography**. [S.l.]: Springer, 1979. p. 379–386.

TOFFOLI, T.; MARGOLUS, N. **Cellular automata machines: a new environment for modeling**. [S.l.]: MIT press, 1987.

TORQUATO, S. A. Cana-de-açúcar para indústria: o quanto vai precisar crescer. **Análises e Indicadores do Agronegócio**, v. 1, n. 10, 2006.

TRIGO, T. R.; JÚNIOR, P. C. d. S. B.; NETO, F. B. D. L. Redes neurais artificiais em colheita de cana-de-açúcar. In: **V Congresso Brasileiro de Agroinformática**. [S.l.: s.n.], 2005.

TSCHIEDEL, M.; FERREIRA, M. F. Introdução à agricultura de precisão: conceitos e vantagens. **Ciência Rural**, Universidade Federal de Santa Maria, v. 32, n. 1, 2002.

VARSHAVSKY, A.; BAKAYEV, V. Studies on chromatin. iv. evidence for a toroidal shape of chromatin subunits. **Molecular biology reports**, v. 2, n. 3, p. 247–254, 1975.

VIANNA, M. **Design thinking: inovação em negócios**. [S.l.]: Design Thinking, 2012.

WEIMAR, J. R. **Simulation with cellular automata**. [S.l.]: Logos-Verlag, 1997.

WILLIAMSON, D. F.; PARKER, R. A.; KENDRICK, J. S. The box plot: a simple visual method to interpret data. **Annals of internal medicine**, Am Coll Physicians, v. 110, n. 11, p. 916–921, 1989.

WOLFRAM, S. **Cellular automata as simple selforganizing systems**. [S.l.], 1982.

_____. Statistical mechanics of cellular automata. **Reviews of modern physics**, APS, v. 55, n. 3, p. 601, 1983.

_____. Universality and complexity in cellular automata. **Physica D: Nonlinear Phenomena**, Elsevier, v. 10, n. 1-2, p. 1–35, 1984.

_____. **A new kind of science**. [S.l.]: Wolfram media Champaign, IL, 2002. v. 5.

YARA. Princípios agronômicos da cana-de-açúcar. **Disponível em: <https://www.yarabrasil.com.br/nutricao-de-plantas/cana-de-acucar/principios-agronomicos-da-cana-de-acucar/>. Acesso em: 26 julho. 2018.**, 2018.

YUE, S.; PILON, P. A comparison of the power of the t test, mann-kendall and bootstrap tests for trend detection/une comparaison de la puissance des tests t de student, de mann-kendall et du bootstrap pour la détection de tendance. **Hydrological Sciences Journal**, Taylor & Francis, v. 49, n. 1, p. 21–37, 2004.

Anexos

ANEXO **A**

Termo de permissão de uso de dados

Recife-PE, 21 de fevereiro de 2018.

Ofício Nº. 003/2018
Do: PPG Informática Aplicada - UFRPE

Ao: Departamento de Agronomia - Usina São José S.A

Prezado(a) Senhor(a) ,

Vimos, através deste ofício, solicitar de V. S.ª permissão para que o discente Geraldo Gomes da Cruz Júnior, CPF 092.038.374-26, utilize dados coletados na Usina São José S.A para confecção de sua Dissertação de Mestrado.

Aproveitamos para deixar explícito que os dados serão usados exclusivamente para fins acadêmicos

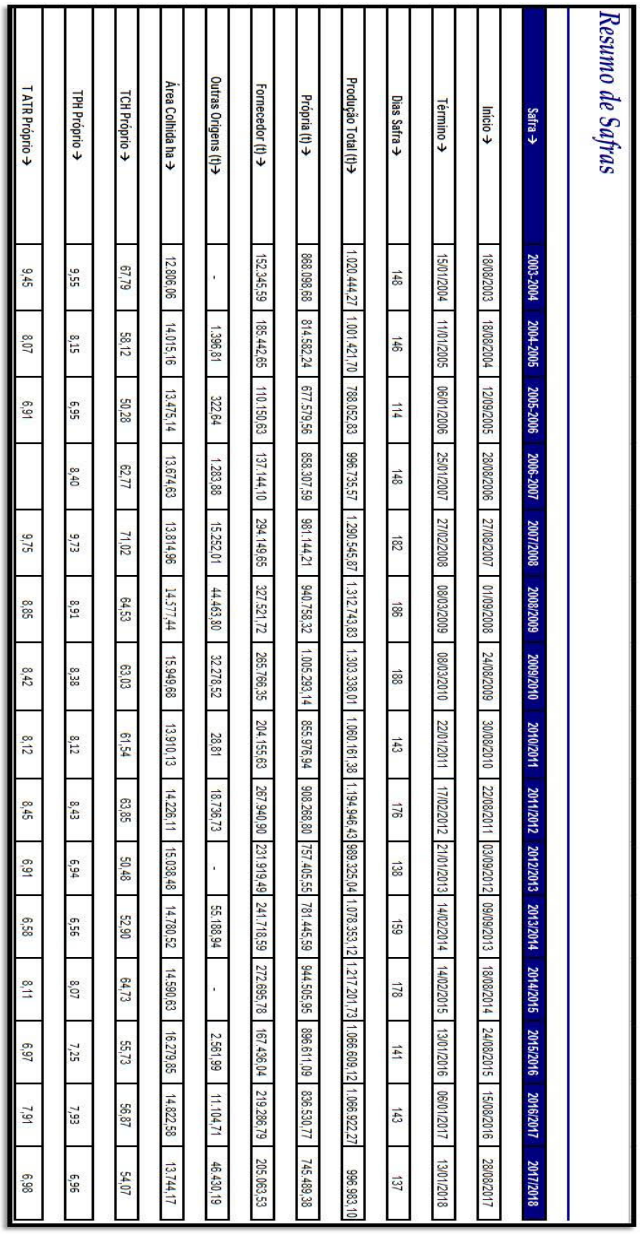
Atenciosamente,

Tiago Alessandro Espínola Ferreira
Coordenador do Programa de Pós-Graduação em Informática Aplicada

Prof. Dr. Tiago Alessandro Espínola Ferreira
Coordenador do Programa de
Pós-Graduação em Informática Aplicada
UFRPE

ANEXO B

Resumo de Safras



Resumo de Safras

Safra →	2003-2004	2004-2005	2005-2006	2006-2007	2007/2008	2008/2009	2009/2010	2010/2011	2011/2012	2012/2013	2013/2014	2014/2015	2015/2016	2016/2017	2017/2018
Início →	19/08/2003	18/08/2004	12/09/2005	28/08/2006	27/08/2007	01/09/2008	24/08/2009	30/08/2010	22/08/2011	03/09/2012	09/09/2013	19/08/2014	24/08/2015	15/08/2016	28/08/2017
Término →	15/01/2004	11/01/2005	06/01/2006	25/01/2007	27/02/2008	08/03/2009	08/03/2010	22/01/2011	17/02/2012	21/01/2013	14/02/2014	14/02/2015	13/01/2016	06/01/2017	13/01/2018
Dias Safra →	148	146	114	148	182	186	188	143	176	138	159	178	141	143	137
Produção Total (t)→	1.020.444,27	1.001.421,70	788.052,83	996.735,57	1.290.545,87	1.312.743,83	1.303.338,01	1.060.161,38	1.194.946,43	689.325,04	1.078.353,12	1.217.201,73	1.066.809,12	1.068.922,27	998.983,10
Própria (t) →	868.098,68	814.582,24	677.579,56	858.307,59	981.144,21	940.758,32	1.005.293,14	855.976,94	908.268,80	757.405,55	781.445,59	944.505,95	896.611,09	836.530,77	745.469,38
Fornecedor (t) →	152.345,59	185.442,65	110.150,83	137.144,10	294.149,65	327.521,72	265.766,35	204.155,63	267.940,90	231.919,49	241.718,59	272.695,78	167.436,04	219.286,79	205.063,53
Outras Origens (t)→	-	1.396,81	322,64	1.283,88	15.252,01	44.463,80	32.278,52	28,81	18.736,73	-	55.188,94	-	2.561,99	11.104,71	46.430,19
Área Colhida ha →	12.806,06	14.015,16	13.475,14	13.674,63	13.814,96	14.577,44	15.949,68	13.910,13	14.226,11	15.038,48	14.780,52	14.590,63	16.279,85	14.822,58	13.744,17
TCH Próprio →	67,79	58,12	50,28	62,77	71,02	64,53	63,03	61,54	63,85	50,48	52,90	64,73	55,73	56,87	54,07
TPH Próprio →	9,55	8,15	6,95	8,40	9,73	8,91	8,38	8,12	8,43	6,94	6,56	8,07	7,25	7,93	6,96
TATR Próprio →	9,45	8,07	6,91		9,75	8,85	8,42	8,12	8,45	6,91	6,58	8,11	6,97	7,91	6,98

Figura 49 – Resumo de Safras