**Wanderley Clarete Lanza Meirelles**

# Estruturação do Problema e Análise de Similaridade Aplicados na Redução do Número de Locais de Experimentos em Ensaios Nacionais de Milho

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador: Prof. Dr. Luis Enrique Zárate Galvez

Belo Horizonte

Abril 2008

FICHA CATALOGRÁFICA

Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Meirelles, Wanderley Clarete Lanza

M514e Estruturação do problema e análise de similaridade, aplicados na redução do número de locais de experimentos em ensaios nacionais de milho / Wanderley Clarete Lanza Meirelles. – Belo Horizonte, 2008.
126 f.

Orientador: Prof. Dr. Luís Enrique Zárate Galvez.
Dissertação (mestrado) – Pontifícia Universidade Católica de Minas Gerais, Programa de Pós-graduação em Educação.
Bibliografia.

1. Milho – Banco de dados. 2. Milho – Melhoramento genético. 3. Redes neurais (Computação) I. Galvez, Luís Enrique Zárate. II. Pontifícia Universidade Católica de Minas Gerais. III. Título

CDU: 681.3.011: 633.15

Bibliotecário: Fernando A. Dias – CRB6/1084

PUC Minas
Programa de Pós-graduação em Informática

# FOLHA DE APROVAÇÃO

"Estruturação do Problema e Análise de Similaridade Aplicados na Redução do Número de Locais de Experimentos em Ensaios Nacionais de Milho"

Wanderley Clarete Lanza Meirelles

Dissertação defendida e aprovada pela seguinte banca examinadora:

Prof. Luis Enrique Zárate Gálvez - Orientador (PUC Minas)
Doutor em Engenharia Metalúrgica e de Minas - UFMG

Prof. José Luis Braga (UFV)
Doutor em Informática – PUC Rio

Prof. Clodoveu Augusto Davis Júnior (PUC Minas)
Doutor em Ciência da Computação – UFMG

Profa. Cristiane Neri Nobre (PUC Minas)
Doutora em Bioinformática - UFMG

Belo Horizonte, 11 de abril de 2008.

Anel Rodoviário – KM 23,5, Rua Walter Ianni, 255 – São Gabriel / Belo Horizonte – MG / CEP: 31980-110
Fone: (31) 3439-5204 – Fax: (31) 3439-5204 – e-mail: mestradoinfsec@pucminas.br

# Dedicatória

Dedico este trabalho à minha esposa Tânia,
à minha filha Beatriz, à minha família e a
todos os meus amigos que, de alguma forma,
me incentivaram a vencer mais este
desafio na minha vida.

# Agradecimentos

Primeiramente à Embrapa, empresa que tenho orgulho de trabalhar e que me proporcionou todas as condições para a realização deste curso.

Ao meu orientador, Professor Luís Enrique Zárate, pela paciência, amizade e pelos inúmeros ensinamentos repassados.

Aos colegas da Embrapa Antônio Carlos, Carla, Cleso, Enilda, Gisela, Luiz Marcelo, Paulo César Magalhães e Paulo Evaristo pela colaboração e pela disposição em ajudar. Sem a ajuda de vocês certamente esta caminhada seria bem mais árdua.

Aos colegas do mestrado Anne, José Geraldo, Leonardo, Michelle e Sandro pela amizade e colaboração, fundamentais para vencer os desafios que enfrentamos durante todo curso.

À Giovana, pelo sorriso sempre presente e pela disposição em ajudar.

A minha esposa pelo constante incentivo e a minha filhinha Beatriz, que, mesmo sem saber falar uma só palavra, muito me ajudou.

# Resumo

O ensaio nacional de milho, conjunto de experimentos científicos realizados em diversas cidades do Brasil, tem como objetivo avaliar o desempenho de cultivares (plantas) de milho, em diferentes condições de solo, clima e altitude. Nesses experimentos, o desenvolvimento da cultura é monitorado durante todo o ciclo do plantio, coletando-se informações acerca do comportamento das plantas. Essas informações são posteriormente utilizadas para selecionar as cultivares mais indicadas para cada região e também para se avaliar a viabilidade do lançamento comercial de cultivares em processo de desenvolvimento. Nesta dissertação, investiga-se a utilização do processo de descoberta de conhecimento em banco de dados (KDD - *Knowledge Discovery in Databases*) aplicado aos dados gerados a partir dos experimentos que compõem o ensaio nacional de milhos híbridos elite, safra 2003/2004, com o objetivo de verificar a possibilidade de se reduzir o número de experimentos realizados, sem perda na qualidade e representatividade dos dados necessários para a perfeita avaliação das cultivares plantadas. Para isso, a interação entre solo, clima e planta de milho foi modelada e os principais componentes de cada um destes sistemas foram identificados. Também é investigada a aplicabilidade do uso de redes neurais artificiais para estimar dados de umidade relativa do ar, necessários ao modelo definido. Utilizando técnicas de *Data Mining* (análise de cluster), foram construídos agrupamentos de experimentos (cidades) que, de acordo com a metodologia proposta neste trabalho, são similares em termos de solo, clima e comportamento das plantas de milho. A coerência dos agrupamentos formados foi avaliada através de dois critérios, sendo o primeiro deles proposto nesta dissertação e o segundo definido por pesquisadores em melhoramento genético da Embrapa Milho Sorgo. Na avaliação pelo primeiro critério, dos 11 agrupamentos formados, quatro foram considerados incoerentes e pelo segundo critério, nenhum agrupamento foi considerado coerente.

# Abstract

The national assay of maize, a set of scientific experiments conducted in several cities in Brazil, aims to evaluate the performance of cultivars of maize (plants) in different conditions of soil, climate and altitude. In these experiments, the development of the culture is monitored during the planting cycle, and information about the plants behavior is collected. This information is later used to select the most suitable cultivars for each region and also to evaluate the viability of the commercial launch of cultivars in the development process. In this dissertation, it is investigated the use of Knowledge Discovery in Databases (KDD) applied to the data generated from experiments that are part of the corn hybrids elite national test, crop 2003/2004, with the main goal of verifying the possibility of reducing the number of experiments carried out, without loss in quality and representativeness of the data needed to perfect the evaluation of the cultivars planted. In order to accomplish this purpose, the interaction between soil, climate and plant maize was modeled and the main components of each one of these systems were identified. It´s also investigated the applicability of artificial neural networks to estimate data on relative humidity, required for the defined model. With the use of Data Mining techniques (cluster analysis), were constructed groupings of experiments (cities) that, according to the methodology proposed in this paper, are similar in terms of soil, climate and behavior of the plants of maize. The consistency of the groups was evaluated by two criteria, the first of them, as proposed in this dissertation and the second, defined by researchers in genetic improvement of Embrapa Milho e Sorgo.

## LISTA DE FIGURAS

## LISTA DE TABELAS

**SUMÁRIO**

# 1. INTRODUÇÃO

O milho, uma das culturas mais antigas do mundo, é talvez a mais importante planta comercial com origem nas Américas. Indícios de seu cultivo encontrados em escavações remontam a 5000 anos atrás. Após o descobrimento da América, foi levado para a Europa e cultivado em jardins, antes mesmo de seu valor alimentício tornar-se conhecido. O cultivo do milho em escala comercial espalhou-se desde a latitude de 58º Norte (União Soviética) até 40º Sul (Argentina). Hoje o milho disputa com o trigo o título de grão mais plantado no mundo.

Pela sua versatilidade, o milho é utilizado desde a alimentação animal, passando pela agricultura de subsistência até a indústria de alta tecnologia. Atualmente, com a crescente preocupação com a geração de combustíveis renováveis, sua importância tem aumentado ainda mais como fonte geradora de energia limpa e renovável.

No Brasil, a utilização do milho na alimentação animal representa, em média, 70% da produção, conforme pode ser visto na Tabela 1.1. O consumo de milho em grão na alimentação humana, apesar de baixo, é fator importante, uma vez que em algumas situações é a fonte de energia diária, sobretudo em regiões de baixa renda, como o semi-árido do Nordeste brasileiro.

**Tabela 1.1**
**Estimativa do consumo de milho no Brasil em 2001**

| Uso | (%) |
|---|---|
| Consumo Animal | 63,5 |
| Avicultura | 32,4 |
| Suinocultura | 20,7 |
| Pecuária | 6,7 |
| Outros Animais | 3,7 |
| Industrial | 10,0 |
| Consumo Humano | 3,6 |
| Perdas e Sementes | 0,6 |
| Exportação | 13,6 |
| Outros | 8,7 |
| Total | 100,0 |

**Fontes: Abimilho, MB Associados e Safras & Mercado**

O Brasil é o terceiro maior produtor mundial de milho, ficando atrás dos Estados Unidos e da China. Entretanto, nossa importante indústria de aves e suínos nos torna um dos maiores consumidores do mundo, fazendo com que seja crescente a necessidade de maior produção de milho.

Apesar de terceiro maior produtor mundial de milho, o Brasil não se destaca entre os países com maior nível de produtividade. Na Tabela 1.2 pode-se observar a produtividade dos maiores produtores de milho do mundo. De acordo com dados da FAO (*Food and Agriculture Organization of the United Nations*), a produtividade média mundial em 2001 estava em torno de 4427 kg/ha, o que nos coloca abaixo desta média.

Vale ressaltar que a cultura do milho no Brasil ainda conta com grandes possibilidades de aumento de produção via crescimento da produtividade, o que tem ocorrido sistematicamente nos último anos.

**Tabela 1.2**
**Produtividade média mundial de milho (kg/ha)**

| Ano | Brasil | USA | China | Argentina | México | França | Romênia | Índia | Itália | Mundo |
|---|---|---|---|---|---|---|---|---|---|---|
| **1990** | 1874 | 7438 | 4525 | 3461 | 1994 | 6019 | 2761 | 1518 | 7638 | 3680 |
| **1991** | 1808 | 6817 | 4580 | 4044 | 2052 | 7277 | 4077 | 1376 | 7262 | 3686 |
| **1992** | 2283 | 8253 | 4535 | 4524 | 2345 | 7964 | 2047 | 1676 | 8660 | 3893 |
| **1993** | 2532 | 6321 | 4964 | 4355 | 2440 | 8045 | 2605 | 1602 | 8664 | 3625 |
| **1994** | 2363 | 8700 | 4695 | 4237 | 2226 | 7792 | 3132 | 1448 | 8225 | 4114 |
| **1995** | 2601 | 7123 | 4918 | 4522 | 2288 | 7717 | 3192 | 1585 | 8970 | 3792 |
| **1996** | 2697 | 7978 | 5204 | 4040 | 2239 | 8382 | 2932 | 1709 | 9336 | 4226 |
| **1997** | 2623 | 7952 | 4390 | 4556 | 2384 | 9059 | 4176 | 1746 | 9627 | 4143 |
| **1998** | 2796 | 8438 | 5269 | 6077 | 2343 | 8453 | 2756 | 1755 | 9346 | 4433 |
| **1999** | 2760 | 8398 | 4946 | 5182 | 2560 | 8901 | 3628 | 1655 | 9744 | 4363 |
| **2000** | 2736 | 8603 | 4670 | 5444 | 2166 | 9058 | 1556 | 1769 | 9386 | 4230 |
| **2001** | 3352 | 8672 | 4933 | 5592 | 2557 | 8593 | 2419 | 1807 | 8942 | 4427 |

**Fonte: FAO, 2002**

O caminho para se chegar a esse aumento de produtividade e, assim, colocar o Brasil no mesmo patamar de países como Estados Unidos, França e Itália passa por muitas frentes. Uma delas, obrigatoriamente, é a pesquisa científica.

O desenvolvimento de cultivares de milho mais produtivas, resistentes às doenças e mais adaptadas às condições brasileiras é condição *sine qua non* para se elevar a produtividade média nacional.

O processo de desenvolvimento de cultivares de milho é longo, envolve um grande número de profissionais (geneticistas, estatísticos, biólogos, fitopatologistas e muitos outros), além de significativo aporte de recursos financeiros. Resultados concretos são obtidos muitas vezes após anos de trabalho e pesquisa.

Quando uma nova cultivar é desenvolvida, esta precisa ser plantada e avaliada juntamente com várias outras para que se avalie seu comportamento e seu desempenho em diferentes ambientes e assim verificar a viabilidade de seu lançamento comercial.

Essa avaliação é feita através de uma rede de ensaios (experimentos) que são implantados em diferentes cidades do país e obedecendo a técnicas próprias de plantio. Em seguida, todo o desenvolvimento da cultura é monitorado e vários parâmetros são avaliados e medidos, gerando um grande volume de dados, que são posteriormente analisados estatisticamente.

Em face da grande extensão territorial do Brasil e da conseqüente diversidade de condições climáticas, de solo e altitude, um número expressivo de ensaios, em diferentes cidades, é necessário para melhor expressar as interações das cultivares avaliadas com o meio ambiente. Quanto maior o número de ensaios instalados, melhor será descrita essa interação. Entretanto, a instalação desses ensaios é onerosa, uma vez que envolve o deslocamento de equipes para o plantio, além do monitoramento constante durante todo o ciclo da cultura. Outro agravante é que, muitas vezes, restrições orçamentárias obrigam à redução da quantidade de ensaios instalados. É nesse contexto que este trabalho se insere e pretende contribuir.

Atualmente o processo de descoberta de conhecimento em banco de dados, também conhecido como KDD (*Knowledge Discovery in Databases*), vem sendo utilizado com sucesso em grandes massas de dados para detectar comportamentos e tendências que esses dados "escondem". Há relatos da utilização bem sucedida desse processo em diversos segmentos, sobretudo em aplicações comerciais e industriais.

Utilizando o processo de KDD, os dados gerados a partir dos ensaios de milho foram analisados com vistas a reduzir a quantidade de locais onde os experimentos são instalados, sem perda na qualidade e confiabilidade nos resultados. Técnicas de mineração de dados, mais especificamente análise de similaridade (*clustering)* foram utilizadas para tal.

## 1.1 Justificativa

Do ponto de vista do agronegócio, os ensaios nacionais de milho são instrumentos de grande importância para agricultores, extensionistas e pesquisadores da área de melhoramento genético, pois fornecem uma grande quantidade de informações acerca das cultivares avaliadas nas diversas cidades e regiões.

A utilização de técnicas de análise de cluster sobre os dados gerados pelos ensaios poderia identificar regiões com características comuns, identificando assim "cidades-típicas" que representassem, com fidelidade, uma região ou um conjunto de cidades, dispensando então a instalação de ensaios nos outros locais com características semelhantes.

A eventual identificação destas "cidades-típicas" representaria um avanço na condução dos ensaios, uma vez que poder-se-ia ter uma redução no número de ensaios instalados e, conseqüentemente, economia de tempo e recursos financeiros. Essa economia poderia ser utilizada na instalação de ensaios em áreas hoje não cobertas pela rede, proporcionando assim informações para agricultores que hoje não se beneficiam dos dados gerados pelos ensaios nacionais.

Do ponto de vista científico e tecnológico, a utilização do processo de KDD em dados de pesquisa agropecuária é pouco relatada na literatura. O estudo e aplicação dessas técnicas, caso seja comprovada sua viabilidade e confiabilidade, daria um novo enfoque ao tratamento dos dados gerados, o que poderia auxiliar e complementar a forma tradicional de se fazer pesquisa agropecuária.

Caso comprovada a viabilidade e confiabilidade das técnicas de clusterização aplicadas aos dados de Ensaios nacionais de milho, a metodologia utilizada poderia ser extrapolada para ensaios nacionais de várias culturas, em diversos centros de pesquisas em todo Brasil.

## 1.2 Objetivos

O objetivo geral do trabalho foi modelar o sistema planta de milho – meio ambiente, levantando seus principais componentes e suas interações. De posse desse modelo, os dados nele definidos foram levantados para serem utilizados como entrada para o processo KDD e

para as técnicas de mineração de dados, mais especificamente os algoritmos de análise de similaridade (*clustering*), a fim de agrupar as diversas cidades onde foram realizados os ensaios em blocos com características semelhantes. O objetivo é tentar reduzir o número de cidades necessárias ao plantio de experimentos, sem perda na qualidade e representatividade dos resultados.

Os objetivos específicos são:

- Através de estudos e entrevistas com especialistas de diversas áreas como solo, clima, fisiologia vegetal, melhoramento genético e outros, definir quais os componentes que influenciam o desenvolvimento de uma planta de milho, definindo assim um modelo (caracterização do problema) com os principais componentes e suas interações.
- Levantar os dados necessários ao modelo e que, em princípio, não fazem parte da base de dados gerada pela rede de ensaios, a fim de enriquecê-la e fornecer subsídios para que a interação planta - meio ambiente possa ser melhor avaliada;
- Aplicar as técnicas de análise de similaridade sobre dados do Ensaio Nacional de Milhos Híbridos Elite[1], safra 2003/2004, a fim de tentar reduzir o número de cidades necessárias à implantação de experimentos, mantendo a confiabilidade dos resultados;
- Validar os resultados encontrados junto aos especialistas de cada área (pesquisadores da Embrapa), analisando a pertinência e coerência dos resultados.

## 1.3 Principais contribuições

As principais contribuições deste trabalho são listadas abaixo:

- Proposta de um modelo da interação planta de milho - solo - clima com a definição dos componentes de cada um destes sistemas;
- Análise e mapeamento das influências e interações das variáveis de solo e clima em cada uma das características observadas nos ensaios nacionais de milho;

[1] Milhos Híbridos Elite são milhos avançados e com grande potencial de lançamento comercial.

- Construção de uma representação utilizando o modelo de redes neurais para estimar a umidade relativa do ar média, a partir de dados de latitude, longitude, altitude, temperatura mínima e temperatura máxima;
- Definição de agrupamentos de cidades, que, de acordo com a metodologia proposta neste trabalho, são consideradas similares em termos de solo, clima e comportamento das plantas, o que permite a redução do número de ensaios instalados no Brasil.

## 1.4 Organização do trabalho

O trabalho trata da aplicação do processo de KDD e da utilização de técnicas e algoritmos de mineração de dados aplicados a um conjunto de dados de experimentos científicos utilizados para avaliar o comportamento de 32 cultivares (plantas) de milho em 35 experimentos (ensaios) realizados em 27 cidades. O trabalho está dividido em capítulos de acordo com os temas pertinentes à sua realização.

No Capítulo 2 é apresentado o Ensaio Nacional de Milho, a fundamentação sobre o processo de descoberta de conhecimento em banco de dados (KDD) além de uma revisão sobre a utilização das técnicas de mineração de dados no setor agrícola. São apresentados ainda os fundamentos das Redes Neurais Artificiais (RNA) e uma revisão sobre sua utilização na previsão de dados climáticos.

No Capítulo 3 é apresentada a metodologia utilizada e todas as etapas percorridas até a montagem do banco de dados.

No Capítulo 4 é apresentado o modelo desenvolvido para estimar o valor médio da umidade relativa do ar, para diferentes cidades do país, através da utilização de RNA.

No Capítulo 5 a etapa de mineração de dados é descrita e os resultados obtidos são analisados

No Capítulo 6 são apresentadas as conclusões e possíveis trabalhos futuros.

# 2. REVISÃO BIBLIOGRÁFICA-CONCEITUAL

## 2.1 O Ensaio nacional de milho

O Ensaio Nacional de Milho é composto por um conjunto de experimentos realizados em diversas cidades, obedecendo aos mesmos procedimentos de plantio. Em cada um dos experimentos são plantadas as mesmas cultivares (plantas) de milho e todo o desenvolvimento da cultura é monitorado. Dados sobre produção de grãos, produção de espigas, altura das plantas, ataque de pragas, doenças e vários outros são coletados, formando assim uma base de comparação entre o desempenho de cada cultivar, nos diferentes locais onde foram realizados os ensaios.

Sua execução é feita de forma cooperada entre instituições públicas e privadas sob a coordenação da Embrapa Milho e Sorgo, unidade da Empresa Brasileira de Pesquisa Agropecuária (Embrapa), com representantes da Associação Brasileira de Sementes e Mudas (ABRASEM). Seus principais objetivos são:

- Avaliar, nas principais regiões produtoras, as cultivares de milho desenvolvidas por entidades públicas e privadas;
- Auxiliar agricultores e técnicos na escolha das cultivares de milho mais adaptadas às suas regiões;
- Regionalizar a recomendação de cultivares de acordo com a altitude, temperatura e tolerância às principais doenças foliares e pragas;
- Fornecer dados para registro de cultivares.

A proteção de cultivares, instituída pela lei n.º 9456, de 25 de abril de 1997, tem por finalidade resguardar os direitos relativos à propriedade intelectual sobre plantas. Os Ensaios Nacionais de Milho, além de fornecerem dados importantes aos melhoristas e auxiliar os técnicos e agricultores na escolha das cultivares mais adaptadas a cada região, são uma importante fonte de informações para cumprimento das exigências legais com vistas ao lançamento e comercialização de cultivares de milho no Brasil.

## 2.2 O processo de descoberta de conhecimento em banco de dados – (KDD - *Knowledge Discovery In Databases*)

As últimas décadas foram marcadas por uma grande e acelerada geração de dados e de informações. Aliada a isso, a crescente utilização de recursos computacionais, advinda principalmente da evolução tecnológica e conseqüente redução de custos, propiciou que estes dados fossem armazenados em meio eletrônico.

Estes fatos contribuíram para que grandes bancos de dados fossem gerados. Empresas, instituições governamentais e centros de pesquisas despendem hoje quantidade razoável de recursos financeiros e humanos a fim de montar bancos de dados sobre os mais variados temas. Mas, por mais paradoxal que possa parecer, grandes bancos de dados não significam necessariamente grandes conhecimentos sobre o tema em questão.

Segundo Kantardzic (2003), apenas uma pequena parte destes dados são efetivamente utilizados porque, na maioria dos casos, os volumes são de tal ordem que o gerenciamento ou a complexidade das estruturas desses dados inviabilizam sua utilização efetiva para análises mais acuradas.

Ainda de acordo com Kantardzic (2003), a razão primária deste problema é que o esforço para a criação dos bancos de dados é freqüentemente direcionado por assuntos como eficiência de armazenamento e desempenho e não por um plano de utilização destes dados. Outro aspecto já apontado por Piatetsky (1991) é a inviabilidade de se analisar estes dados, produzidos e armazenados em larga escala, através de métodos manuais tradicionais, tais como planilhas de cálculos e relatórios.

Diante deste cenário, surgiu a necessidade de explorar estes dados para extrair informação e conhecimento implícito, neles contidos. Extrair conhecimento novo, potencialmente útil e que, em princípio, está oculto em grandes massas de dados é a essência do processo denominado KDD (*Knowledge Discovery in Databases*).

Segundo Fayyad *et al.* (1996), o processo KDD refere-se às etapas que produzem conhecimentos a partir dos dados e, principalmente, à etapa de mineração dos dados, que é a fase que pode transformar dados em conhecimento e informação. Esse processo consiste em encontrar e interpretar padrões nos dados, de modo iterativo e interativo, através da intervenção de um agente humano no processo KDD e da repetição dos algoritmos e da análise de seus resultados.

Segundo (Witten e Frank, 2000), “mineração de dados é definida como o processo de descoberta de padrões nos dados. Esse processo precisa ser automático ou semi-automático e os padrões descobertos precisam ser significativos e ter algum retorno econômico, além de serem necessários, para tal processo, dados em quantidade substancial”.

Em sua base, KDD é a união de conhecimentos de banco de dados, estatística, inteligência artificial e aprendizado de máquina (*machine learning*). Utilizando técnicas avançadas, como redes neurais, indução de regras, árvores de decisão, análise de agrupamentos e muitas outras, KDD vem sendo cada vez mais utilizado para análises em grandes massas de dados e para detectar comportamentos e tendências que esses dados “escondem”.

O processo de KDD compreende diversas etapas, não necessariamente seqüenciais, pois em cada uma delas pode-se necessitar voltar a etapas anteriores. A Figura 2.1 mostra as etapas do processo KDD, segundo Fayyad *et al*. (1996). Esta abordagem define a iteratividade das etapas e a interatividade com o usuário no processo. Ao final de cada etapa, os resultados obtidos são analisados para que, com base na experiência e conhecimento do usuário eventuais correções sejam efetuadas.

O termo *Data Mining* (mineração de dados) é freqüentemente relacionado ao processo de KDD, mas *Data Mining* é apenas uma das etapas do processo. Nesta etapa diversas técnicas podem ser utilizadas, mas nos concentraremos na análise de agrupamentos, também conhecida como “clusterização”, que foi objeto de estudo neste trabalho.

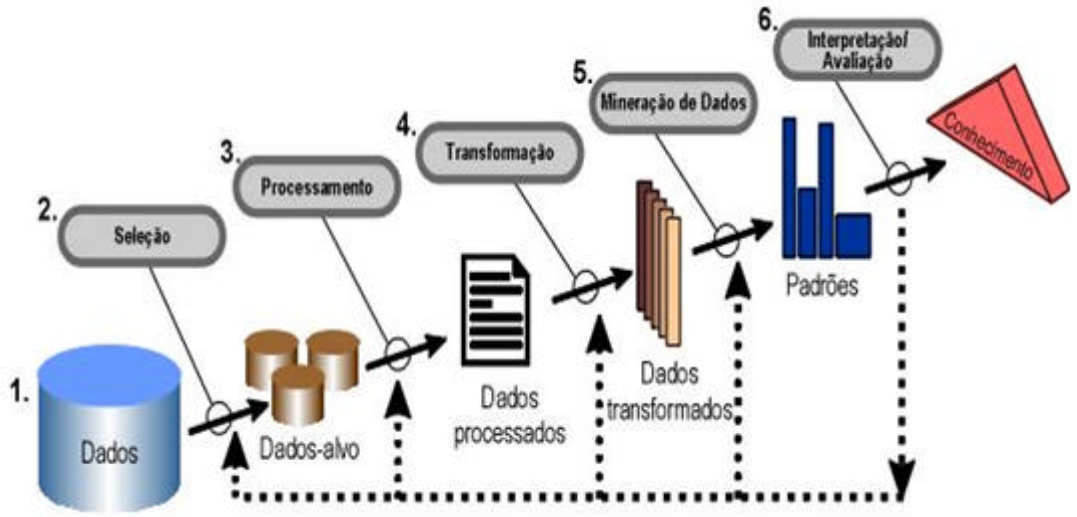


Figura 2.1 – Etapas do processo de KDD segundo Fayyad *et al*. (1996)

### 2.2.1 Etapas do processo KDD

#### 2.2.1.1 Definição do espaço problema

A primeira tarefa a ser realizada é uma profunda análise do problema a ser resolvido. É de vital importância o perfeito entendimento do domínio da aplicação para que se possa definir corretamente os objetivos do processo de KDD. Só assim é possível verificar se o conhecimento descoberto é de fato útil.

#### 2.2.1.2 Seleção de atributos

Com base no entendimento do domínio, seleciona m-se os atributos que são relevantes para se obter o resultado esperado. O sucesso do processo KDD depende da escolha correta destes dados, pois é neste conjunto que serão aplicados os algoritmos para descoberta de conhecimento.

#### 2.2.1.3 Preparação e pré-processamento da base de dados

Usualmente os dados selecionados para o processo de KDD não estão em um formato adequado para a extração de conhecimento. Além disso, é necessário assegurar a qualidade desses dados. À adequação desses dados é feita através das etapas de limpeza, transformação, enriquecimento e melhoramento. Essas etapas e as ações nelas executadas são conhecidas como pré-processamento.

De acordo com Kantardzic (2003), as principais tarefas realizadas durante o pré-processamento dos dados são:

1. Detecção e remoção de *outliers*

   *Outliers* são dados com valores incomuns ou incompatíveis com a maioria das observações de um determinado atributo. Essas amostras podem afetar gravemente os resultados obtidos na mineração de dados.

2. Ampliação/redução e codificação dos dados

   O pré-processamento dos dados envolve diversos passos, como adequação das escalas das variáveis, redução da dimensionalidade da base de dados, codificação de atributos de forma a possibilitar uma melhor interpretação pelos algoritmos de mineração de dados.

Ainda de acordo com Kantardzic (2003), o pré-processamento dos dados não deve ser considerado uma etapa completamente independente das demais, uma vez que, a cada iteração do processo de mineração de dados, alterações na base de dados podem ser definidas para melhorar os resultados nas iterações seguintes.

De acordo com Pyle (1999), o pré-processamento dos dados consome 75% do tempo gasto em um processo de KDD. A seções seguintes apresentam estas etapas.

#### 2.2.1.3.1 Limpeza

Os dados selecionados deverão ser filtrados, para que inconsistências e ruídos (repetição de tuplas, tipos incompatíveis, etc.) sejam identificados e removidos. O preenchimento ou eliminação de valores nulos também é feito nesta etapa, que consome a maior parte do tempo em um processo KDD. Ao final da etapa de limpeza, a base de dados deve estar corrigida e consistente.

#### 2.2.1.3.2 Enriquecimento

Nesta fase são agregados aos dados existentes, informações que possam contribuir no processo de descoberta do conhecimento. Essas informações não estão na base de dados, mas

são conhecidas e acredita-se que podem melhorar o desempenho da etapa de mineração de dados. Esses dados são definidos na estruturação do problema.

#### 2.2.1.3.3 Melhoramento

Melhoramento dos dados consiste em realçar características dos dados sem, no entanto adicionar dados de fontes externas. O melhoramento pode ser através da extração de características adicionais de um determinado atributo ou através da variação de sua amostragem.

#### 2.2.1.3.4 Transformação ou codificação

A padronização dos dados para que a mesma informação não esteja armazenada de forma diferente (ex. produção em toneladas ou quilos), a transformação de valores contínuos de um atributo em um número pré-determinado de intervalos (discretização) são algumas das tarefas efetuadas nesta etapa. Nesta fase os dados são convertidos para a forma mais adequada para a utilização nos algoritmos de mineração.

### 2.2.1.4 Mineração de dados

É a etapa mais importante do processo de KDD e é caracterizada pela busca de padrões nos dados. Nesta etapa, é escolhido o método e são definidos os algoritmos que realizarão a busca pelo conhecimento implícito e útil do banco de dados. Essa escolha é feita de acordo com os objetivos definidos anteriormente. As principais técnicas utilizadas nesta etapa são: árvores de decisão, regras de associação, análises de agrupamento, redes neurais, algoritmos genéticos e lógica nebulosa, dentre outras.

#### 2.2.1.5 Avaliação / interpretação dos resultados

É a última etapa do processo de KDD, onde é realizada a interpretação dos resultados obtidos após a aplicação dos algoritmos de *Data Mining*. As saídas obtidas devem ser avaliadas, refinadas e validadas através de medidas de qualidade e através da percepção de especialistas do domínio em estudo.

Considerando que os resultados da mineração de dados devem ajudar na tomada de decisões, é desejável que os modelos obtidos sejam de fácil interpretação. No entanto, Kantardzic (2003) destaca que modelos mais simples e de fácil interpretação usualmente são menos precisos.

Os resultados do processo de descoberta do conhecimento podem ser mostrados de forma consolidada em relatórios e tabelas a fim de facilitar a sua interpretação. Após a análise e validação criteriosa dos resultados, deve-se identificar a necessidade de retornar a qualquer uma das etapas anteriores do processo de KDD, caso os resultados não sejam satisfatórios.

### 2.2.2 Análise de similaridade *(Clustering)*

Análise de similaridade, análise de cluster ou análise de agrupamento é um conjunto de metodologias para reconhecimento não supervisionado de padrões, cujo objetivo é separar amostras semelhantes em grupos conhecidos como cluster. As amostras dentro de um cluster são mais similares entre si e menos similares (dissimilares) com as amostras de outro cluster. Everitt (1980), diz que, análise de cluster é um nome genérico para um conjunto de técnicas que produz uma classificação a partir de um conjunto de dados não classificados previamente. O conjunto de amostras a agrupar é representado como um vetor de medidas ou, mais formalmente, como um ponto em um espaço multidimensional.

Para o ser humano é relativamente fácil agrupar amostras em um espaço de uma, duas ou até três dimensões, mas a maioria dos problemas envolve um espaço de muitas dimensões, o que torna inviável realizar esta tarefa manualmente, sem a utilização das técnicas e algoritmos de *clustering*.

O grau de semelhança entre as amostras é dado por uma medida de proximidade. Esta medida é que quantifica o grau de semelhança (similaridade) ou diferença existente entre duas

amostras.

O critério de agrupamento é o conjunto de funções ou regras estabelecidas para a formação dos grupos ou cluster. A medida de similaridade e o critério de agrupamento constituem pontos-chave no processo de *clustering*.

Uma entrada em um processo de análise de cluster pode ser descrita como um par *(X,s)* onde X é um conjunto de amostras e *s* é uma medida de similaridade ou distância (diferença) entre as amostras. Como saída têm-se um conjunto de agrupamentos ? = {$G_1$, $G_2$,... $G_n$} onde $G_k$, com k = 1,...,n é um subconjunto de X onde:

$$G_1 \cup G_2 \cup ... \cup G_N = X \text{ , e}$$

$$G_i \cap G_j = f, \forall\ i \neq j$$

Em princípio, o problema de agrupar amostras pode parecer simples, mas a variedade de formas e tamanhos que os grupos podem apresentar em um espaço n-dimensional o torna um problema difícil e complexo. Outro fator que contribui para aumentar essa complexidade é a quantidade de grupos a formar, pois este número é em princípio desconhecido e depende de fatores como a dimensionalidade dos dados.

Há uma grande disponibilidade de algoritmos e ambientes de software para a tarefa de análise de agrupamentos. Esses algoritmos são divididos em grupos e sub-grupos, a saber:

- Algoritmos de agregação;
- Algoritmos seqüenciais;
- Algoritmos hierárquicos;
  - Algoritmos de aglomeração;
  - Algoritmos de divisão;
- Algoritmos baseados em função de otimização.

Segundo Zaiane (2003), um bom método de análise de cluster deve apresentar as seguintes características:

- Ser capaz de lidar com dados com muitas dimensões (muitos atributos), ou seja, um bom método de análise de cluster deve trabalhar de forma eficiente com conjuntos de baixa ou alta dimensão;

- Ser flexível com a quantidade de elementos a agrupar; ou seja, um bom método de análise de agrupamento deve funcionar de forma eficiente com conjuntos de qualquer tamanho;
- Lidar com diferentes tipos de dados (atributos numéricos, categóricos, nominais, binários);
- Ser capaz de definir clusters de diferentes tamanhos e formas;
- Exigir o mínimo conhecimento para determinação de parâmetros de entrada para o algoritmo de agrupamento;
- Robustez - ser capaz de lidar com ruídos e inconsistências na base de dados;
- Apresentar resultados consistentes, não importando a ordem em que os dados são apresentados;

Segundo Haldiki (2001), nenhum algoritmo atende a todos os requisitos listados acima. Por isso, a escolha do algoritmo mais adequado para cada caso depende de um profundo conhecimento de cada um deles, de modo a selecionar o mais adequado a cada caso. É importante ressaltar que não há uma técnica de *clustering* que é universalmente aplicável a qualquer estrutura de dados multidimensional.

#### 2.2.2.1 Medidas de similaridade

Como dito anteriormente, a medida de similaridade quantifica o grau de semelhança (similaridade) ou diferença existente entre duas amostras. O conceito de medida de similaridade é fundamental na definição de agrupamentos e é amplamente utilizado na maioria dos algoritmos de cluster (Kantardzic, 2003). Ainda de acordo com Kantardzic (2003), a escolha da medida de similaridade deve ser criteriosa, pois, a qualidade do processo de *clustering* depende desta decisão.

É comum a utilização de uma medida de dissimilaridade ou distância ao invés da medida de similaridade. Essas medidas são normalmente expressas como uma função com propriedades métricas e que apresentam as propriedades descritas abaixo, independentemente de como são calculadas:

Seja *X* um conjunto de observações e *x, y, z* amostras de *X* tais que:

$$x, y, z \hat{I} X.$$

A distância de *x* em relação a *y* é dada por *d(x,y)* de forma que:

1. $d(x,y) > 0 \ \forall \ x \ ? \ y$
2. $d(x,y) = 0 \ \ se \ \ x = y$
3. $d(x,y) = d(y,x)$
4. $d(x,y) = d(x,z) + d(z,y)$

Existem formas diferentes para se calcular a similaridade ou distância para cada categoria de atributos (binários, nominais, categóricos, ordinais, etc.). Este trabalho se ateve ao cálculo de distâncias para atributos numéricos (contínuos).

#### 2.2.2.1.1 Distância euclidiana

De acordo com (Everitt, 1980; Kantardzic, 2003), a distância Euclidiana é a mais conhecida e mais utilizada medida de distância para um espaço multi-dimensional.
Seja $X = [x_{1,} x_{2,...}x_n]$ e $Y = [y_{1,} y_{2,...}y_n]$, a distância euclidiana *d(x,y)* é dada por:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

#### 2.2.2.1.2 Distância de Manhattan

Outra medida utilizada é a distância de Manhattan, também conhecida como distância *city block* ou métrica L1. Ela é definida da seguinte forma:

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + ... + |x_n - y_n| = \sum_{i=1}^{n} |x_i - y_i|$$

A distância de Manhattan às vezes apresenta resultado semelhante à distância euclidiana, porém o efeito da diferença entre duas variáveis é minimizado uma vez que ela não é elevada ao quadrado.

#### 2.2.2.1.3 Distância de Minkowski

A distância de Minkowski é dada por:

$$d(x,y) = \left( \sum_{i=1}^{n} (x_i - y_i)^p \right)^{\frac{1}{p}}$$

A distância Euclidiana e a distância de Manhattan são casos especiais da distância de Minkowski onde $p$ = 2 e 1, respectivamente.

### 2.2.2.2 Matriz de similaridade

A forma mais usual de apresentação da distância ou similaridade entre um conjunto de amostras é através de uma matriz de similaridade. Independente da medida de similaridade ou distância adotada, essa matriz é simétrica, conforme mostrado abaixo:

$$d(i,j) = \begin{bmatrix} d_{12} & d_{13} & \cdots & d_{1n} \\ & d_{23} & \cdots & d_{2n} \\ & & \cdots & \cdots \\ & & \cdots & d_{n-1,n} \end{bmatrix}$$

onde *d(i,j)* representa a distância ou a similaridade entre as amostras *i* e *j (i = 1,...n-1); (j = 2,...,n).*

#### 2.2.2.3 Normalização

As unidades de medida das variáveis observadas podem influenciar no processo de agrupamento, pois variáveis em diferentes unidades ou escalas numéricas podem contribuir de forma desequilibrada no cálculo da similaridade. A fim de solucionar este problema, são propostas diversas técnicas para normalizar ou padronizar os valores das variáveis utilizadas.

Neste trabalho optou-se por normalizar os valores através do mapeamento dos limites inferiores e superiores de cada variável, aplicando-se em seguida a equação 2.1, que sempre resulta em valores entre 0 e 1.

$$Zi = \frac{(Xi - X_{\min})}{(X_{\max} - X_{\min})} \qquad (2.1)$$

Sendo:

$Zi$: valor normalizado da variável $Xi$

$Xi$: valor original da variável $Xi$

$X_{min}$: valor mínimo dentre todos os elementos da variável $X$

$X_{max}$: valor máximo dentre todos os elementos da variável $X$

#### 2.2.2.4 Técnicas de agrupamento

Várias são as técnicas de agrupamento descritas na literatura, sendo duas as abordagens principais: métodos hierárquicos e métodos não hierárquicos ou de particionamento.

Métodos hierárquicos consistem de sucessivas divisões ou agrupamentos de elementos a cada iteração do algoritmo. São subdivididos em métodos aglomerativos e divisivos.

Nos métodos aglomerativos, inicialmente, cada elemento representa um grupo ou cluster, e a cada passo, um novo elemento é anexado a um grupo já formado. Ao final, um único agrupamento contendo todos os elementos é formado. Fazem parte desta abordagem os seguintes métodos:

- Método de ligação por vizinho mais próximo (*Single Linkage*);
- Método de ligação por vizinho mais distante (*Complete Linkage*);
- Método de ligação por média (*Average Linkage*);
- Método de ligação por mediana (*Median Linkage);*
- Método de ligação por centroide (*Centroid Linkage);*
- Método de ligação de Ward .

Nos métodos divisivos, o processo é o inverso, ou seja, inicialmente têm-se apenas um grupo contendo todos os elementos. A cada passo esse grupo subdivide-se em um ou mais grupos até o final do processo, quando se têm cada elemento representando um grupo.

De acordo com Everitt (1980), o primeiro passo de um algoritmo do método divisivo é dividir o grupo inicial em dois grupos. A partir daí estes dois grupos podem ser divididos de $2^{n-1} – 1$ maneiras. A quantidade de possibilidades de divisão cresce de forma exponencial com o número de elementos do grupo inicial. O esforço computacional demandado por esses algoritmos limita sua utilização à bases de dados menores.

Uma implementação deste método que não considera todas as possibilidades de divisão dos grupos é feita pelo algoritmo de *MacNaughton-Smith.*

Nos métodos de agrupamento hierárquico, a cada iteração do algoritmo um elemento é anexado (aglomerativos) ou retirado (divisivos) do cluster. Nos métodos aglomerativos, o algoritmo finaliza quando é formado um único cluster contendo todos os elementos. Já nos métodos divisivos, o algoritmo finaliza quando são formados *n clusters*, cada um contendo somente um elemento. A decisão de quando parar é que determina a quantidade de agrupamentos formados. Essa decisão é subjetiva e deve levar em conta os objetivos da análise.

Nos métodos não hierárquicos ou de particionamento, o número de agrupamentos é definido previamente. Inicialmente são formados agrupamentos na quantidade desejada. A partir daí são calculados os centros de cada grupo e os elementos são realocados para o grupo cujo centro é mais próximo. O processo continua até que se encontre uma estabilidade entre os grupos.

De acordo com Kantardzic (2003), os métodos de particionamento apresentam melhor desempenho que os métodos hierárquicos, principalmente em grandes bases de dados. Em geral o que diferencia os métodos de particionamento é a forma utilizada para definir a melhor partição. Os métodos de particionamento mais conhecidos são o *K-means* e *K- mediods.*

O método *K-means*, utilizado neste trabalho, é detalhado adiante. Descrições detalhadas dos demais métodos aqui citados são encontradas em Bussab *et al*. (1990), Everitt (1980), Kantardzic (2003).

#### 2.2.2.4.1 Método *K-Means*

De acordo com Kantardzic (2003), o *K-means*, também conhecido como K-médias, é um algoritmo de particionamento simples, sendo também o mais utilizado quando se deseja separar um conjunto de dados em uma quantidade de grupos previamente determinada.

O método inicia com grupos aleatórios e vai rearranjando esses grupos baseado na similaridade dos elementos com cada grupo, até que um critério de convergência seja alcançado. Usualmente esse critério é quando não há mais alterações entre os grupos.

A distância euclidiana é comumente utilizada pelo método *K-means* como medida de dissimilaridade entre os elementos a se agrupar.

Seja $E = (e_1,e_2,...e_n)$ um vetor de características, $M = \{ E_{(1)}, E_{(2)},...E_{(m)} \}$, uma matriz contendo *m* amostras de *E*, *k* a quantidade de grupos que se deseja formar então *K-means* tem como entrada *(M,k)* e como saída, um conjunto de *k* grupos de *E* e uma matriz (*k* x *n*) com as médias ou centróides de cada um dos *k* grupos formados, para cada uma das *n* características.

De acordo com Jain *et al*. (1999), o algoritmo *K-means* tem complexidade de tempo de $O(nkl)$ e complexidade de espaço da ordem de $O(k + n)$ sendo *n* o número de elementos, *k* o número de grupos e *l* o número de iterações.

A Figura 2.2 mostra o algoritmo *K-means*.

1. selecionar inicialmente K agrupamentos contendo elementos escolhidos aleatoriamente e calcular os centróides de cada agrupamento;
2. agregar a cada agrupamento o elemento mais próximo de seu centróide;
3. recalcular os centróides dos novos agrupamentos;
4. repetir os passos 2 e 3 até que não ocorra mudanças de elementos entre agrupamentos.

Figura 2.2 – Algoritmo *K-means*

O exemplo abaixo ilustra como *K-means* agruparia o conjunto de dados da Tabela 2.1, em dois grupos *(k = 2)*, usando a distância euclidiana como medida de similaridade.

**Tabela 2.1**
**Conjunto de dados de exemplo *(k-means)***

| Elemento | Variáveis Observadas | |
|---|---|---|
| | X | Y |
| 1 | 12 | 9 |
| 2 | 6 | 21 |
| 3 | 12 | 21 |
| 4 | 6 | 9 |
| 5 | 9 | 15 |
| 6 | 18 | 3 |

- Passo 1
  Definir os *k* agrupamentos iniciais aleatoriamente. Neste caso, G1 = { 1, 2, 3 } e G2 = { 4, 5, 6 }

- Passo 2
  Calcular os centróides de cada grupo

  Centróide G1: $\frac{12+6+12}{3}=10\ (X)$ $\frac{9+21+21}{3}=17\ (Y)$

  Centróide G2: $\frac{6+9+18}{3}=11\ (X)$ $\frac{9+15+3}{3}=9\ (Y)$

- Passo 3
  Calcular a distância euclidiana entre cada elemento e os centróides dos dois grupos.

| Distância entre cada elemento e o centróide do grupo | | |
|---|---|---|
| **Elemento** | **G1** | **G2** |
| 1 | $\sqrt{(12-10)^2+(9-17)^2}=8{,}25$ | $\sqrt{(12-11)^2+(9-9)^2}=1$ |
| 2 | $\sqrt{(6-10)^2+(21-17)^2}=5{,}66$ | $\sqrt{(6-11)^2+(21-9)^2}=13$ |
| 3 | $\sqrt{(12-10)^2+(21-17)^2}=4{,}47$ | $\sqrt{(12-11)^2+(21-9)^2}=12{,}04$ |
| 4 | $\sqrt{(6-10)^2+(9-17)^2}=8{,}94$ | $\sqrt{(6-11)^2+(9-9)^2}=5$ |
| 5 | $\sqrt{(9-10)^2+(15-17)^2}=2{,}24$ | $\sqrt{(9-11)^2+(15-9)^2}=6{,}32$ |
| 6 | $\sqrt{(18-10)^2+(3-17)^2}=16{,}12$ | $\sqrt{(18-11)^2+(3-9)^2}=9{,}22$ |

- Passo 4

  Rearranjar os elementos em cada grupo de acordo com a menor distância entre os elementos e o centróide. Na tabela acima se observa que o elemento 1 está mais próximo do grupo G2 e o elemento 5 está mais próximo do grupo G1, portanto os grupos devem ficar da seguinte forma: G1 = { 2 , 3, 5 } e G2 = {1, 4, 6 }

- Passo 5

  Recalcular os centróides dos novos grupos

  Centróide G1: $\frac{6+12+9}{3} = 9\ (X)$ $\frac{21+21+15}{3} = 19\ \ (Y)$

  Centróide G2: $\frac{12+6+18}{3} = 12\ \ (X)$ $\frac{9+9+3}{3} = 7\ \ (Y)$

- Passo 6

  Recalcular a distância euclidiana entre cada elemento e os centróides dos dois grupos.

| **Distância entre cada elemento e o centróide do grupo** | | |
|---|---|---|
| **Elemento** | **G1** | **G2** |
| 1 | $\sqrt{(12-9)^2+(9-19)^2} = 10{,}44$ | $\sqrt{(12-12)^2+(9-7)^2} = 2$ |
| 2 | $\sqrt{(6-9)^2+(21-19)^2} = 3{,}61$ | $\sqrt{(6-12)^2+(21-7)^2} = 15{,}23$ |
| 3 | $\sqrt{(12-9)^2+(21-19)^2} = 3{,}61$ | $\sqrt{(12-12)^2+(21-7)^2} = 14$ |
| 4 | $\sqrt{(6-9)^2+(9-19)^2} = 10{,}44$ | $\sqrt{(6-12)^2+(9-7)^2} = 6{,}32$ |
| 5 | $\sqrt{(9-9)^2+(15-19)^2} = 4$ | $\sqrt{(9-12)^2+(15-7)^2} = 8{,}54$ |
| 6 | $\sqrt{(18-9)^2+(3-19)^2} = 18{,}36$ | $\sqrt{(18-12)^2+(3-7)^2} = 7{,}21$ |

- Passo 7

  Rearranjar os elementos em cada grupo de acordo com a menor distância entre os elementos e o centróide. Neste ponto, observa-se na tabela acima que os elementos já estão separados nos grupos corretos, uma vez que, a menor distância de cada elemento é exatamente com o centróide do grupo ao qual ele já faz parte. Assim o algoritmo é finalizado e os grupos formados são G1 = { 2, 3, 5} e G2 = { 1, 4, 6 }.

De acordo com Kantardzic (2003), o método *K-means* é simples e computacionalmente eficiente, porém, é muito sensível a ruídos e *outliers* na base de dados

uma vez que, uma pequena quantidade de dados pode influenciar substancialmente o valor dos centróides.

### 2.2.3 A mineração de dados no setor agrícola

Apesar de KDD e mineração de dados serem relatados em trabalhos de diversas áreas do conhecimento, Abdullah *et al.* (2004), dizem que "aplicações de *data mining* tem, muito recentemente, encontrado seu caminho em pesquisa agropecuária".

Técnicas de análise de agrupamento são relatadas em trabalhos cujo objetivo é selecionar ou agrupar plantas com características semelhantes, como pode ser visto em Dias Filho *et al.* (1994), que utilizaram agrupamento hierárquico (vizinho mais distante) e não hierárquico (*k-means*) na avaliação da adaptação agronômica[2] de 118 variedades de uma gramínea forrageira chamada *Panicum Maximum Jacq.* Já Santos *et al*. (2007) utilizaram o método de agrupamento UPGMA (*Unweighted Pair Group Method with Arithmetic Average)* para selecionar genótipos de soja resistentes a ferrugem asiática. Mineração de dados através da utilização de algoritmos genéticos foram utilizados por Guimarães (2005) para avaliar a correlação entre os elementos físico-químicos de água e solo.

Mineração de dados também é relatada em trabalhos relacionados ao estudo de fenômenos climáticos, como pode ser visto em Bucene *et al.* (2002) que utilizaram técnicas *data mining* em dados climáticos para previsão de geada e deficiência hídrica para as culturas do café e da cana-de-açúcar para o estado de São Paulo. Técnicas de agrupamento hierárquico foram utilizadas por Keller Filho *et al.* (2005) para determinar regiões pluviometricamente homogêneas no Brasil. Abdullah *et al.* (2004) também utilizaram análise de agrupamento em dados meteorológicos, juntamente com dados sobre utilização de pesticidas e ataque de pragas para otimizar o uso de pesticidas em culturas de algodão, no Paquistão.

A utilização de técnicas de mineração de dados é freqüentemente relatada em trabalhos relacionados à classificação de imagens de satélite, conforme pode ser visto em Song *et al.* (2005), que utilizaram árvores de decisão, mais especificamente o algoritmo C4.5, para classificar imagens de satélite quanto à cobertura e utilização do solo.

[2] Adaptação agronômica é uma medida integrada relacionada à suscetibilidade a pragas e doenças, vigor de crescimento e recuperação em relação à herbivoridade simulada.

Outra utilização de análise de agrupamento é relatada em Melo *et al.* (2004) que utilizaram o método de ligação de *Ward* para comparar indicadores de produção de soja, em diversos municípios do Rio Grande do Sul, com o zoneamento agrícola.

Outro campo onde a mineração de dados tem se apresentado como poderosa ferramenta é na biologia molecular e prospecção gênica, conforme pode ser visto em Xavier *et al.* (2005) assim como em Emygdio (2003), que utilizaram o método de agrupamento UPGMA para analisar a variabilidade genética por meio de marcadores moleculares RAPD, de feijão caupi e de variedades comerciais de feijão, respectivamente.

Outras aplicações de mineração de dados na agricultura podem ser vistas em Bertis *et al*. (2001), Cunningham and Holmes (2001), Harms *et al*. (2001), Scherte (2002), Yang *et al*. (1999).

### 2.2.4 Redes neurais artificiais - RNA

Segundo (Braga *et al*., 2003; Haykin, 2001; Kantardzic, 2003), Redes Neurais Artificiais são sistemas paralelos distribuídos por unidades de processamento simples (nodos) que calculam determinadas funções matemáticas, normalmente não lineares. Essas unidades são dispostas em uma ou mais camadas interligadas por um grande número de conexões que por sua vez são associadas a pesos, os quais são ajustados para encontrar a representação neural do processo. As RNA têm a habilidade de aprender a partir da experiência e seu funcionamento é inspirado no cérebro humano.

No contexto da inteligência artificial, RNA têm sido cada vez mais estudadas e utilizadas, principalmente pela sua capacidade de representar problemas não lineares por aprendizado e pela sua capacidade de generalização.

A Figura 2.3 mostra o modelo de um neurônio artificial, base para o projeto das RNA, de acordo com Haykin (2001). Nela podem-se identificar os elementos de um neurônio artificial, conforme descrito abaixo:

1. Sinais de entrada ($x_1..x_m$);
2. Sinapses ou elos de conexão, caracterizados pelos pesos sinápticos ($w_{k1}..w_{km}$);
3. Um somador (? ) responsável por totalizar os sinais de entrada ponderados pelos pesos sinápticos;

4. Um bias ($b_k$) cuja função é aumentar ou reduzir a entrada líquida da função de ativação;
5. Uma função de ativação [f(.)] para restringir a amplitude da saída de um neurônio;
6. Um valor de saída ($y_k$)

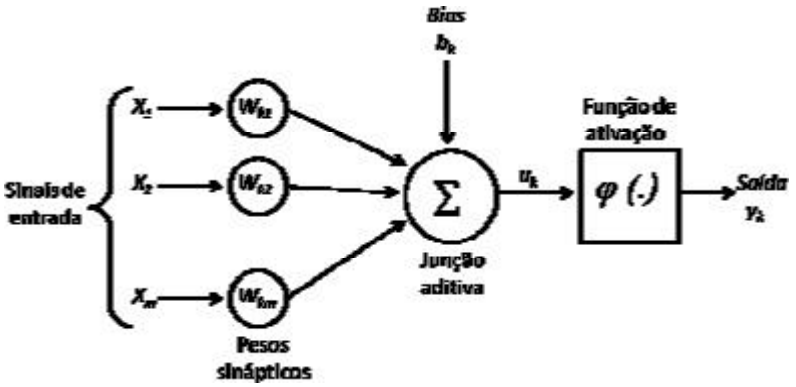


Figura 2.3 – Modelo de neurônio artificial segundo Haykin (2001)

Matematicamente o neurônio *k* pode ser descrito pelas equações abaixo:

$$u_k = \sum_{j=1}^{m} w_{kj} x_j$$

$$y_k = j\,(u_k + b_k)$$

Onde:

$x_1, x_2, .., x_m$ são sinais de entrada;

$w_{k1}, w_{k2}, ..., w_{km}$ são os pesos sinápticos do neurônio *k*;

$u_k$ é a saída do somador

$b_k$ é o bias

f(.) é a função de ativação

$y_k$ é o sinal de saída do neurônio.

Uma rede neural artificial é formada por um conjunto de neurônios artificiais simples, onde cada neurônio executa uma atividade simples, mas a rede neural completa possui capacidade computacional para resolver problemas complexos.

Na utilização de RNA, na primeira fase, conhecida como treinamento, um conjunto de exemplos do problema que se deseja representar é apresentado à rede, a qual aprende a relação entre esses exemplos. Segundo Kantardzic (2003), uma RNA aprende sobre seu

ambiente com um processo iterativo de ajustes aplicados aos pesos das conexões. Idealmente, a rede torna-se mais "conhecedora" sobre seu ambiente após cada iteração no processo de aprendizagem.

O poder das RNA vai muito além do simples mapeamento entre dados de entrada e saída, pois sua capacidade de generalização confere à rede a habilidade de dar respostas coerentes para dados (entradas) não vistas na fase de treinamento. De acordo com Braga *et al.* (2003), as RNA são capazes de atuar como mapeadores universais de funções multivariáveis, com custo computacional que cresce apenas linearmente com o número de variáveis.

No processo de aprendizagem ou treinamento de uma RNA, usualmente é empregado um método conhecido como "aprendizado supervisionado". Neste método são fornecidos à rede, por um agente externo, conjuntos de dados de entrada e os resultados esperados (saída), com o objetivo de ajustar os parâmetros da rede de forma a encontrar a relação entre os pares entrada e saída. Assim a rede tem sua saída calculada comparada com a saída desejada. O erro encontrado é utilizado para ajustar os pesos das conexões com vistas a minimizá-lo. Este processo se repete para cada padrão apresentado à rede. A redução do erro, ou seja, da diferença entre a saída calculada pela rede e a saída esperada é incremental, pois ajustes nos pesos são feitos a cada etapa de treinamento de tal forma que se caminhe para a diminuição do erro. A rede é considerada treinada quando o erro atinge um patamar aceitável e previamente estabelecido ou depois de certo número de iterações.

Existem diversos algoritmos de treinamento de RNA, a maioria deles do tipo supervisionado, sendo o *back-propagation* (retro-propagação do erro) (Rumelhart e McClelland, 1986) o mais conhecido deles. Nele, o treinamento ocorre em duas fases conhecidas como fase *forward* e fase *backward*. Na fase *forward* a saída da rede é calculada para um dado padrão e na fase *backward* os pesos das conexões são ajustados de acordo com a saída desejada e a saída calculada pela fase *forward*. Diversos algoritmos de otimização têm sido aplicados em conjunto com o algoritmo *back-propagation* visando determinar o conjunto de pesos que gere o menor erro com maior taxa de convergência. Dentre esses algoritmos destacam-se o do gradiente descendente com *momentum*, também conhecido como regra delta generalizada ou algoritmo *back-propagation* tradicional e o Levenberg-Marquardt, sendo este considerado o método mais rápido para treinamento de redes neurais (Barbosa *et al.*, 2005).

### 2.2.4.1 RNA Multicamadas

Uma RNA multicamadas, também conhecida de *Multilayer Perceptron* (MLP) é o tipo mais importante e utilizado de RNA (Kantardzic, 2003). Uma rede MLP típica é constituída por um conjunto de neurônios representando a camada de entrada, um conjunto de neurônios distribuídos em uma ou mais camadas intermediárias (camada oculta) e outro conjunto de neurônios representando a camada de saída da rede. A Figura 2.4 mostra uma rede MLP com cinco entradas ($X_1..X_5$), uma camada oculta com quatro neurônios e a camada de saída com dois neurônios ($Y_1$,$Y_2$). O sinal de entrada se propaga da esquerda para direita, camada por camada. O processamento realizado por um neurônio é definido pela combinação dos processamentos realizados pelos neurônios da camada anterior.

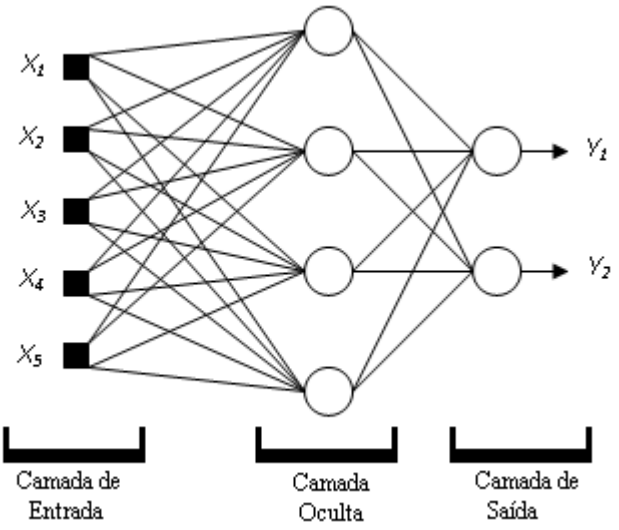


Figura 2.4 – Uma rede MLP totalmente conectada

Uma rede MLP possui a capacidade de tratar dados que não são linearmente separáveis. Isto confere às redes MLP um poder computacional muito maior que uma RNA de uma só camada. Segundo Haykin (2001), as redes MLP têm sido utilizadas com sucesso para resolver problemas de difícil solução, através de seu treinamento utilizando o algoritmo *back-propagation*.

Ainda de acordo com Haykin (2001), uma rede MLP apresenta três características distintas:

1. O modelo de cada neurônio da rede inclui uma função de ativação não-linear. Essa não linearidade deve ser suave, isto é, diferenciável em qualquer ponto. Uma função normalmente utilizada e que satisfaz esta exigência é uma não-linearidade sigmóide definida por $y_j = \frac{1}{1+\exp(-v_j)}$.
2. A rede contém uma ou mais camadas de neurônios ocultos que não são parte da entrada nem da saída da rede. Estes nodos ocultos permitem que a rede aprenda tarefas complexas extraindo progressivamente as características mais significativas dos padrões de entrada.
3. A rede possui um alto grau de conectividade entre as camadas.

Através das combinações destas características juntamente com a habilidade de aprender através do treinamento, é que a rede MLP deriva seu poder computacional.

A utilização de redes neurais é amplamente descrita na literatura, principalmente em aplicações de reconhecimento, classificação e recuperação de padrões, previsão de séries temporais, aproximação de funções, dentre outros.

#### 2.2.4.2 Redes neurais para previsão de dados climáticos

Atualmente redes neurais artificiais vêm sendo amplamente utilizadas na representação de processos físicos. O uso de RNA na área de climatologia tem sido investigado por diversos autores (Silva *et al.,* 2006; McCullagh *et al.*, 1999; Pessoa *et al.*, 2006). De acordo com Silva *et al.* (2006), redes neurais apresentam-se como ferramenta adequada à previsão agrometeorológica. Pela sua capacidade de aprendizado e generalização, redes neurais artificiais são cada vez mais utilizadas neste tipo de tarefa, conforme poder ser visto em McCullagh *et al*. (1999), citados por Bucene *et al.* (2002), que utilizaram técnicas de redes neurais artificiais para estimar parâmetros meteorológicos como a precipitação. Já Zanetti *et al.* (2005) utilizou RNA para estimar a evapotranspiração[3] de referência. Pessoa *et al.* (2006) também utilizaram RNA em simulações climáticas para a América do Sul visando determinar anomalias de precipitação e temperatura sazonal.

[3] Evapotranspiração é considerada como a perda de água por evaporação do solo e transpiração das plantas

Neste trabalho, RNA multicamadas foram utilizadas para estimar dados ausentes de umidade relativa do ar. O modelo proposto e todas as etapas do processo utilizado são descritos no capítulo 4.

# 3. METODOLOGIA DE DESENVOLVIMENTO

Neste capítulo é descrita a utilização do processo KDD com vistas à redução da quantidade de ensaios de milho. Todas as etapas que antecedem a etapa de mineração de dados são descritas de forma detalhada. O problema objeto de estudo é modelado e todos os seus componentes definidos. A seguir todas as variáveis necessárias são levantadas e trabalhadas para que, ao final, o banco de dados esteja pronto para ser utilizado na etapa de mineração de dados.

Na etapa de preparação da base de dados, foi constatada a ausência de informações climáticas consideradas pelos especialistas consultados como imprescindíveis. Foi utilizado um modelo computacional neural para estimar estas informações. A fim de facilitar o entendimento da metodologia utilizada para a simulação destes dados, optou-se por descrevê-la em um capítulo a parte (Capítulo 4). A etapa de mineração de dados e análise dos resultados será descrita no Capítulo 5.

## 3.1 Definição do escopo do problema

A montagem de ensaios para avaliação de cultivares segue regras de estatística experimental. São montados em diversas cidades, sempre com as mesmas cultivares, visando assim criar uma base de comparação para se avaliar o comportamento de cada cultivar em determinado local.

Ao fim do ciclo do plantio e com todos os dados coletados, são geradas planilhas para cada cidade, conforme esquema abaixo:

Cidade AA

| | Característica A | Característica B | .... | Característica M |
|---|---|---|---|---|
| Cultivar 1 | | | | |
| Cultivar 2 | | | | |
| .... | | | | |
| Cultivar N | | | | |

No ano agrícola 2003/2004, objeto de estudo neste trabalho, foram realizados 35 experimentos em 30 cidades e coletadas informações sobre 11 características de 32 cultivares.

A Tabela 3.1 mostra as cidades onde foram realizados os experimentos, separadas por unidade da federação. A Figura 3.1 mostra o mapa com essas cidades.

**Tabela 3.1**
**Relação de cidades onde foram realizados ensaios safra 2003/2004**

| UF | Cidade |
|---|---|
| ES | Sooretama |
| GO | Cristalina, Goianésia, Goiânia, Ipameri, Itumbiara, Montividiu, Morrinhos, Palmeiras de Goiás, Planaltina, Porangatu, Rio Verde |
| MA | São Raimundo das Mangabeiras |
| MG | Brasilândia de Minas, Patos de Minas, Sete Lagoas, Uberlândia |
| MS | Dourados, Maracaju, Ponta Porã |
| MT | Sinop |
| PI | Baixa Grande Ribeiro |
| PR | Brasilândia do Sul, Campo Mourão, Londrina, Palotina |
| RO | Vilhena |
| SP | Assis, Birigui , Piracicaba |

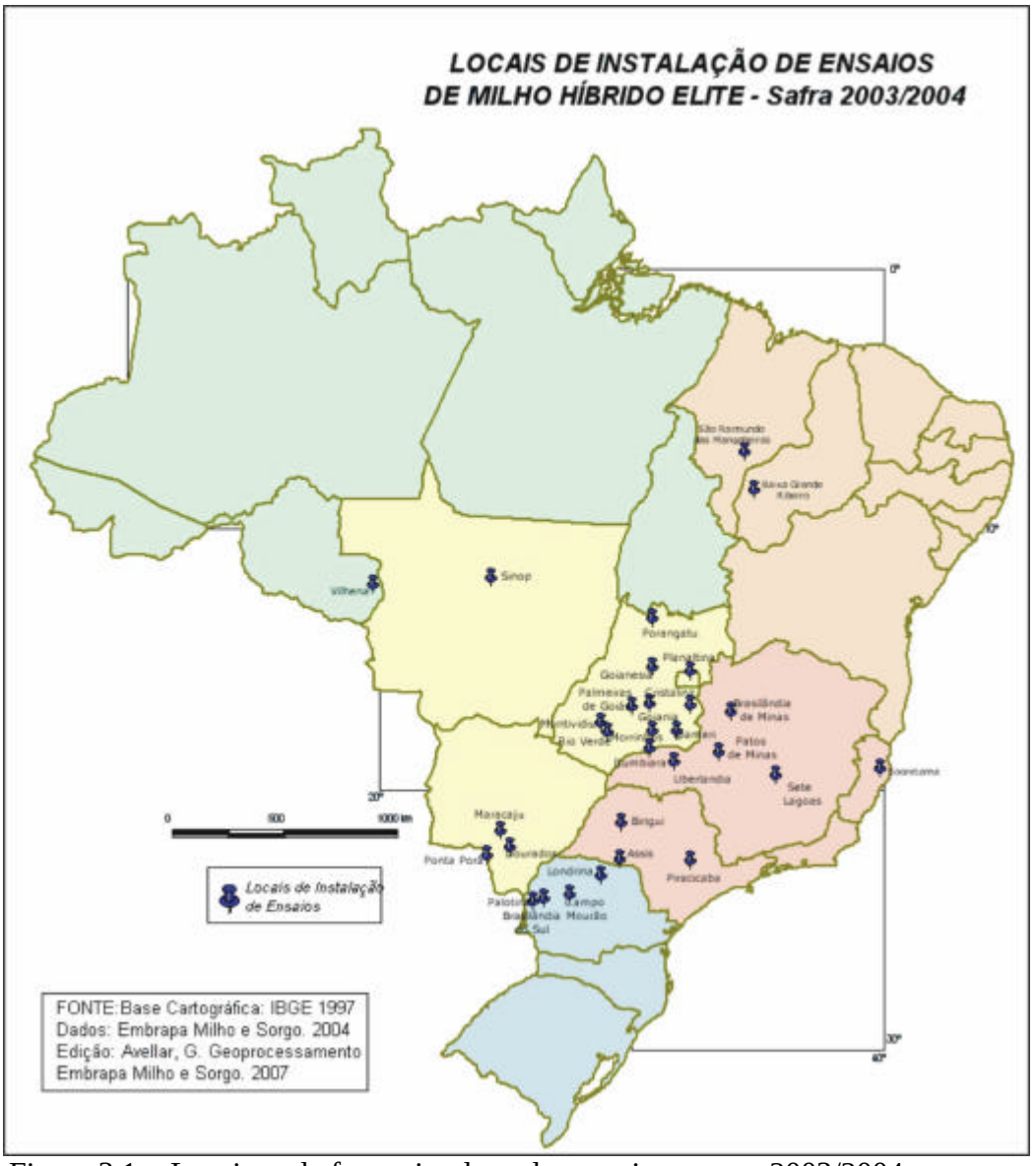


Figura 3.1 – Locais onde foram implantados ensaios no ano 2003/2004.

Nas cidades de Campo Mourão, Goiânia, Londrina, Maracaju, Palotina, Ponta Porã, Rio Verde e Sete Lagoas foram realizados dois ensaios, sendo um deles com plantio nos meses de fevereiro e março. Esses ensaios são chamados de safrinha, pois, são plantados em datas posteriores à safra normal, que é plantada entre outubro e dezembro.

A Tabela 3.2 mostra as características avaliadas em cada experimento.

**Tabela 3.2**
**Características avaliadas nos ensaios 2003/2004**

| Variável | Descrição |
|---|---|
| Cultivar | Nome da cultivar |
| Peso de Espigas (kg/ha) | Peso das espigas ajustados para kg/ha, com o grau de umidade corrigido para 13%, a partir da pesagem das espigas colhidas na área |
| Peso de Grãos (kg/ha) | Peso de grãos ajustados para kg/ha, com o grau de umidade corrigido para 13%, a partir da pesagem dos grãos de todas as espigas colhidas na área |
| Florescimento (dias) | Período em dias, decorrido da emergência ao florescimento feminino (emissão dos estilo-estigmas) em 50% das plantas da parcela |
| Altura da Planta (cm) | Altura da planta medida em cm, da superfície do solo à inserção do pendão com a folha bandeira |
| Altura da Espiga (cm) | Altura da espiga, medida em cm, da superfície do solo à inserção da espiga superior |
| Acamadas + Quebradas (%) | Percentual de plantas acamadas e quebradas[4] |
| Estande final | Estande final determinado pelo número de plantas por parcela |
| Número de espigas | Número total de espigas na parcela |
| Espigas doentes (%) | Percentual de espigas doentes na parcela |
| Umidade (%) | Umidade média dos grãos na parcela |

## 3.2 Estruturação do domínio do problema

Em princípio, os dados disponíveis se limitavam às planilhas com os dados coletados acerca dos experimentos. Nessas planilhas somente são anotados dados observados sobre o plantio. Estes dados dizem respeito tão somente ao desenvolvimento das plantas.

Como é sabido, uma planta, assim como qualquer ser vivo, é o resultado da interação de sua carga genética ou genótipo com o meio ambiente, ou seja, uma planta é o resultado da interação de três fatores: genótipo, solo e clima. Assim sendo, levantou-se através da literatura e de entrevistas com especialistas em solo, clima, fisiologia vegetal e melhoramento genético, quais seriam os componentes que influenciam o desenvolvimento de uma planta de milho.

[4] Plantas acamadas são plantas que, na colheita, apresentam ângulo de inclinação igual ou inferior a 45° em relação ao solo. Plantas quebradas são plantas que, na colheita, apresentam colmo quebrado abaixo da inserção da espiga.

Para este levantamento foi considerada também a disponibilidade das informações. Sobre o solo, chegou-se às variáveis relacionadas na Tabela 3.3. A altitude foi relacionada junto às variáveis de solo. Sobre o clima, chegou-se às variáveis listadas na Tabela 3.4. Em relação às plantas e sua carga genética, optou-se por não relacionar nenhum componente, uma vez que, a utilização de informações em nível de genes tornaria este trabalho um problema complexo a ser tratado.

**Tabela 3.3**
**Relação de variáveis sobre o solo**

| Variável | Descrição |
|---|---|
| Saturação por bases (V%) | Saturação por bases (V%) - A capacidade de troca de cátions de Al, H, Ca, Mg e K que um solo é capaz de reter é chamada de CTC. A saturação por bases (V%) é definida como a proporção da CTC, a pH 7, formada por bases. Quanto maior o valor de V, mais fértil é o solo. No tangente à produtividade, a saturação por bases entre 50% e 60% é considerada adequada à maioria dos cereais e entre 60% e 70% adequada à maioria das leguminosas. |
| Textura | A textura do solo refere-se à proporção relativa em que se encontram, em determinada massa de solo, os diferentes tamanhos de partículas. Refere-se, especificamente, às proporções relativas das partículas ou frações de areia, silte e argila. É muito importante na irrigação porque tem influência direta na taxa de infiltração de água, na aeração, na capacidade de retenção de água, na nutrição, como também na aderência ou força de coesão nas partículas do solo. Os solos são agrupados em três classes de textura: Arenosa, Média e Argilosa. Solos de textura arenosa (solos leves) possuem teores de areia superiores a 70% e de argila inferiores a 15%. Solos de textura Média (solos médios) apresentam certo equilíbrio entre os teores de areia, argila e silte. Solos de textura argilosa apresentam teores de argila superiores a 35% |
| Água disponível | Fração da água presente no solo que se encontra em condições de ser absorvida pelas raízes da planta. Em geral, é considerada como o teor de água retida entre a capacidade de campo *in situ* e o ponto de murcha permanente. A quantidade de água disponível varia em função de diferentes atributos do solo e da planta e de condições meteorológicas. |
| Altitude | É a distância vertical de um ponto que o separa do nível médio do mar. |

**Tabela 3.4**
**Relação de variáveis consideradas sobre o clima**

| Variável | Descrição |
|---|---|
| Temperatura máxima | É a maior temperatura registrada em um determinado ponto do espaço. |
| Temperatura mínima | É a menor temperatura registrada em um determinado ponto do espaço. |
| Precipitação | Precipitação atmosférica em forma de gotas de água (chuva), medida através de pluviômetros e expressada através de um valor em mm. |
| Vento | Movimento do ar, geralmente de uma área de pressão mais alta para uma área de pressão mais baixa. |
| Radiação solar | Conjunto de radiações emitidas pelo Sol de curto comprimento de onda (entre 0,15 e 4 mm) e que correspondem a cerca de 99% da radiação solar que atinge a Terra. É dada em número de horas de sol incidentes durante um dia. |
| Umidade relativa do ar | Relação entre a quantidade de vapor d'água contida no ar e a quantidade máxima que o ar pode conter sob as mesmas condições de temperatura e pressão. É expressa como uma porcentagem |

De posse dessas informações e visando tornar mais claras as influências e interações das variáveis de solo e clima em cada uma das características observadas nos ensaios, montou-se o mapa cognitivo mostrado na Figura 3.2. No mapa as setas indicam quais são as características afetadas por uma determinada variável. Através deste mapa chegou-se ao quadro de influências mostrado na Tabela 3.5. Têm-se assim a interação entre planta, solo e clima modelada e também a relação das variáveis que deverão ser levantadas para a montagem do banco de dados.

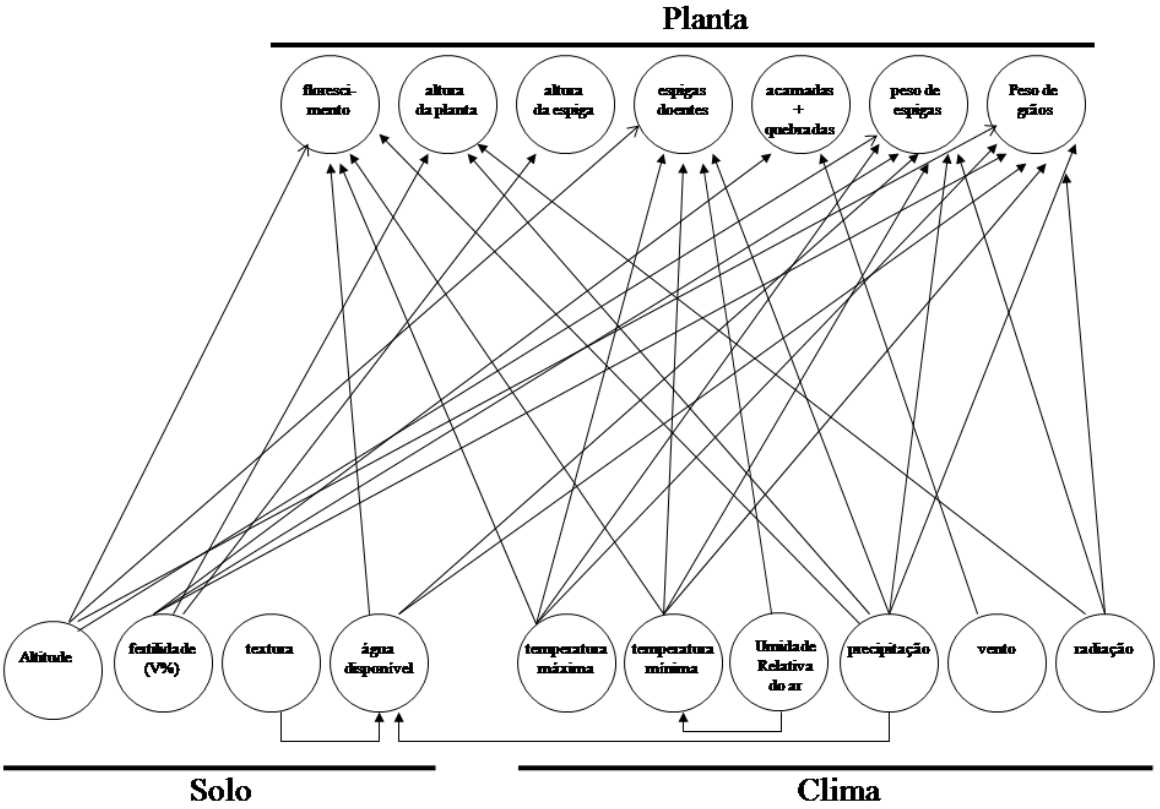


Figura 3.2 – Mapa cognitivo das influências do solo e clima nas variáveis observadas nos ensaios

Em relação às variáveis de clima, foi necessário definir a freqüência mais adequada para tratamento destes dados (diários, semanais, quinzenais ou mensais).

O ciclo do plantio de milho varia de 120 a 180 dias. A utilização de dados diários geraria um grande volume de dados, o que aumentaria o esforço computacional demandado pelos algoritmos de análise de cluster, podendo inviabilizar sua utilização. Em entrevistas com especialistas em fisiologia vegetal e agrometeorologia, ficou estabelecido que a variação diária dos dados climáticos não afetaria o desenvolvimento da planta, e que estes dados poderiam ser utilizados de forma agregada.

O desenvolvimento da planta de milho é divido em fases denominadas estádios. Esses estádios são bem conhecidos e com características próprias de necessidades de nutrientes, água e temperatura. Assim, foram definidos pelos especialistas em fisiologia vegetal os estádios mais importantes onde deveria se coletar os dados de clima, chegando-se aos sete estádios listados na Tabela 3.6.

A mudança destes estádios é definida através de características nas plantas tais como, quantidade de folhas, aparecimento dos estilos-estigmas (cabelo do milho), dentre outras. Como estes dados não são observados durante os experimentos, estimou-se a quantidade

média de dias em que cada um deles acontece, a partir da data de plantio.

A Tabela 3.6 mostra as características principais de cada um dos sete estádios selecionados, assim como o tempo médio em que ocorrem. Informações detalhadas sobre todos os estágios podem ser obtidas em http://www.cnpms.embrapa.br/publicacoes/milho/ecofisiologia.htm.

**Tabela 3.5**
**Influências de planta - solo - clima**

| Características | Influenciado por: |
|---|---|
| Florescimento | Água disponível<br>Temperatura máxima<br>Temperatura mínima<br>Altitude<br>Precipitação |
| Altura da planta | Saturação por bases (v%)<br>Precipitação<br>Radiação |
| Altura da espiga | Saturação por bases (V%) |
| Espigas doentes | Temperatura máxima<br>Temperatura mínima<br>Altitude<br>Umidade relativa |
| Acamadas + quebradas | Saturação por bases (v%)<br>Vento |
| Peso de espigas | Saturação por bases (v%)<br>Água disponível<br>Temperatura máxima<br>Temperatura mínima<br>Altitude<br>Precipitação<br>Radiação |
| Peso de grãos | Saturação por bases (v%)<br>Água disponível<br>Temperatura máxima<br>Temperatura mínima<br>Altitude<br>Precipitação<br>Radiação |
| Água disponível | Textura<br>Precipitação |
| Temperatura mínima | Altitude |

**Tabela 3.6**
**Fases ou estádios considerados para coleta de dados climáticos**

| Fase | Características | Dias após plantio |
|---|---|---|
| VE | germinação e emergência | 5 |
| V3 | três folhas completamente desenvolvidas | 12 |
| V12 | Perda de duas ou quatro folhas basais. Fase mais crítica para a produção | 45 |
| Florescimento | aparecimento do pendão seguido dos cabelos | 55 |
| Enchimento 1 (R2) | grãos brancos na aparência externa | 65 |
| Enchimento 2 (R4) | grão pastoso | 90 |
| Enchimento 3 (R5) | formação do dente | 100-120 |

**Fonte: http://www.cnpms.embrapa.br/publicacoes/milho/ecofisiologia.htm**

A estrutura de atributos do modelo da interação Planta - Solo - Clima (IPSC) pode ser formalmente definida como:

$$IPSC = \{Pl, So, Cl\}$$

onde:

*IPSC* = interação planta, solo, clima;

*Pl = {p1, p2, ... p7}* sendo as características da planta, onde:

*p1* - florescimento (Fl);

*p2* - altura da planta (AP);

*p3* - altura da espiga (AE);

*p4* - espigas doentes (ED);

*p5* - plantas acamadas + quebradas (Ac);

*p6* - peso de espigas (PE);

*p7* - peso de grãos (PG).

*So = {s1, s2, ... s4}* sendo as características do solo, onde:

*s1* – saturação por bases (V%);

*s2* - textura (Te);

*s3* - água disponível (AD);

*s4* - altitude (Al).

$$Cl = \begin{bmatrix} c_1^1 & c_1^2 & \cdots & c_1^6 \\ c_2^1 & c_2^2 & \cdots & c_2^6 \\ \vdots & \vdots & \vdots & \vdots \\ c_7^1 & c_7^2 & \cdots & c_7^6 \end{bmatrix}$$ sendo as características do clima em sete estádios, onde:

$c_q^r$ , q = 1..7, r = 1,...,6 são respectivamente:

$c_q^1$ - Temperatura máxima (Tmax);

$c_q^2$ - Temperatura mínima (Tmin);

$c_q^3$ - Precipitação (Pr);

$c_q^4$ - Vento (Ve);

$c_q^5$ - Radiação solar (RS);

$c_q^6$ - Umidade relativa (UR).

A estrutura de dados de IPSC pode então ser resumida da seguinte forma:

$$IPSC = \{Pl_1^1 Pl_1^2 ... Pl_1^7 So_1^1 So_1^2 ... So_1^4 Cl_1^1 Cl_1^2 ... Cl_1^6 Cl_2^1 ... Cl_7^1 ... Cl_7^6\}$$

## 3.3 Preparação da base de dados

Após definidas as variáveis necessárias, o próximo passo correspondeu a seleção destas a fim de montar o banco de dados no formato adequado para ser utilizado na etapa de mineração de dados. Optou-se por selecionar e efetuar todo o pré-processamento dos dados de cada tema (dados de ensaio, clima e solo) separadamente, para depois montar o banco de dados no formato requerido pelos algoritmos de análise de cluster.

### 3.3.1 Dados de ensaios

Após análise visual dos dados coletados de ensaios, foi constatada a ausência de dados para diversas cidades. De acordo com os pesquisadores responsáveis pelos ensaios, isso se deveu a dificuldades operacionais para coleta dos mesmos no decorrer do plantio. A Tabela 3.7 mostra, em cada cidade, quais foram os dados coletados.

**Tabela 3.7**
**Relação de variáveis coletadas em cada cidade**

| Cidade | Fl | AP | AE | Ac% | ED% | PE | PG |
|---|---|---|---|---|---|---|---|
| Assis Safrinha-SP | | X | X | X | | X | X |
| Baixo Grande Ribeiro | X | X | X | X | X | X | X |
| Birigui | X | X | X | X | | | X |
| Brasilândia do Sul | | X | X | | | | X |
| Campo Mourão | | X | X | X | X | X | X |
| Campo Mourão (safrinha) | X | | | | | | X |
| Cristalina | | X | X | X | | | X |
| Dourados (safrinha) | | X | X | X | | X | X |
| Goianésia | X | X | X | X | | | X |
| Goiânia | X | X | X | X | X | X | |
| Goiânia (safrinha) | | X | X | X | | X | X |
| Ipameri | X | X | X | X | X | X | X |
| Itumbiara | X | X | X | X | | | X |
| Londrina | X | X | X | X | | | X |
| Londrina (safrinha) | X | X | X | X | | | X |
| Maracaju | | X | X | | | | X |
| Maracaju (safrinha) | | X | X | X | | | X |
| Montividiu (safrinha) | | X | X | | X | X | X |
| Morrinhos | | X | X | X | X | X | |
| Palmeiras de Goiás | | | | | X | X | |
| Palotina | | X | X | X | | | X |
| Palotina (safrinha) | | X | X | X | | | X |
| Patos de Minas | | X | X | X | X | X | X |
| Piracicaba | X | X | X | X | X | X | X |
| Planaltina | X | X | X | X | X | X | X |
| Ponta Porã | | X | X | X | | X | X |
| Ponta Porã (safrinha) | X | X | X | X | | X | X |
| Porangatu | X | X | X | X | X | X | X |
| Rio Verde | | X | X | X | X | X | |
| Rio Verde (safrinha) | | X | X | X | | X | X |
| São Raimundo das Mangabeiras | X | X | X | X | X | X | X |
| Sete Lagoas | | X | X | X | X | X | |
| Sete Lagoas (cerrado) | | X | X | X | X | X | |
| Sooretama | | X | X | X | X | X | |
| Uberlândia | | X | X | X | X | X | |
| Total de ensaios onde foi coletado | 14 | 33 | 33 | 30 | 17 | 23 | 27 |
| % coletado em relação a quantidade total de ensaios | 40% | 94% | 94% | 86% | 49% | 66% | 77% |

Conforme pode ser visto na Tabela 3.7, os dados sobre florescimento foram coletados em apenas 40% das cidades, portanto optou-se por não utilizá-los. Quanto aos demais dados ausentes, optou-se por estimá-los, conforme descrito a seguir, na etapa de preparação dos dados.

#### 3.3.1.1 Preparação dos dados de ensaios

Conforme já demonstrado, o algoritmo *K-means* calcula a distância entre cada uma das variáveis avaliadas para compor a distância total entre cada elemento (tupla), no caso, entre cada cidade onde se realizou o ensaio. Desta forma é necessário que as variáveis avaliadas estejam presentes em todas as cidades, caso contrário, a distância entre uma variável coletada e uma não coletada causaria distorções nos resultados.

Para estimar os valores das variáveis altura de planta (AP), altura de espiga (AE), % acamamento (Ac) e % de espigas doentes (ED), calculou-se a média de cada cultivar, nos locais onde a variável foi coletada e esse valor foi utilizado para as cidades onde não houve coleta da variável, conforme exemplo a seguir:

Substituir valores ausentes da variável AP (coletada em 33 ensaios conforme Tabela 3.7)

Dada altura de planta da cultivar *"k"* no ensaio *"c"*, $AP_k^c$, onde *c = 1 ... 33, k = 1 ... 32* então:

1. Soma de todos os valores coletados de AP para a cultivar "k";

$$S_k = \sum_{c=1}^{33} AP_k$$

2. Cálculo da média de AP ( $\overline{M_k}$ ) para cultivar "k";

$$\overline{M_k} = \frac{S_k}{33}$$

3. Substituir valor ausente de AP da cultivar "k" pela média $\overline{M_k}$ em todos os locais onde AP não foi coletado;

   *Se* $AP_k^c$ é ausente então $AP_k^c = \overline{M_k}$

4. Repetir a partir do passo 1 para todas as 32 cultivares.

De forma análoga, os valores ausentes de AE, Ac, ED foram substituídos.

Para as variáveis peso de espiga (PE) e peso de grão (PG), observa-se na Tabela 3.7 que foram coletadas em 66% e 77% dos ensaios respectivamente e ambas foram coletados apenas em 15 ensaios (43%). Observa-se também que sempre existe a ocorrência de PE ou PG. Neste caso optou-se por estimar os valores ausentes utilizando percentual médio de sabugo, calculado através dos ensaios onde PE e PG foram coletados, da seguinte forma:

Dado o peso da espiga PE da cultivar *"k"* do ensaio *"c"* $PE_k^c$;

Dado o peso de grão PG da cultivar *"k"* do ensaio *"c"* $PG_k^c$;

1. Cálculo do percentual do peso de sabugo para cada uma das 32 cultivares, nos 15 ensaios onde PE e PG foram coletadas, através da equação 3.1:

$$p\_S_k^c = \frac{(PE_k^c - PG_k^c) \times 100}{PE_k^c} \quad (\%) \qquad \text{para } c = 1...15,\ k = 1 \ldots 32 \qquad (3.1)$$

2. Cálculo da média do percentual de sabugo para cada uma das 32 cultivares;

$$\overline{mp\_S_k} = \frac{\sum_{c=1}^{15} p\_S_k^c}{15} \qquad \text{para } k = 1 \ldots 32 \qquad (3.2)$$

De posse do valor médio do percentual de sabugo de cada uma das 32 cultivares, as equações abaixo foram utilizadas para estimar valores ausentes de PE e PG:

$$PE\_est_k^c = \left(\frac{PG_k^c \times 100}{100 - \overline{mp\_S_k}}\right) \qquad \text{para } c = 1..35,\ k = 1 \ldots 32 \qquad (3.3)$$

$$PG\_est_k^c = \frac{(100 - \overline{mp\_S_k}) \times PE_k^c}{100} \qquad \text{para } c = 1..35,\ k = 1 \ldots 32 \qquad (3.4)$$

3. Substituir valor ausente de PE por PE estimado ( $PE\_est_k^c$ );

   *Se* $PE_k^c$ é ausente então $PE_k^c = PE\_est_k^c$

4. Substituir valor ausente de PG por PG estimado ( $PG\_est_k^c$ );

   *Se* $PG_k^c$ é ausente então $PG_k^c = PG\_est_k^c$

Desta forma, cada valor ausente de PE foi substituído pelo valor de PG acrescido da média do percentual de sabugo da respectiva cultivar. Já os valores ausentes de PG foram substituídos pelos valores de PE subtraídos do percentual médio de sabugo das respectivas cultivares.

Tem-se assim, a base de dados sobre os ensaios com valores ausentes preenchidos. A Tabela 3.8 mostra a relação das variáveis presentes e simuladas.

**Tabela 3.8**
**Relação de variáveis sobre ensaios.**

| Cidade | AP | AE | Ac% | ED% | PE | PG |
|---|---|---|---|---|---|---|
| Assis Safrinha-SP | X | X | X | simulado | X | X |
| Baixo Grande Ribeiro | X | X | X | X | X | X |
| Birigui | X | X | X | simulado | Simulado | X |
| Brasilândia do Sul | X | X | simulado | simulado | simulado | X |
| Campo Mourão | X | X | X | X | X | X |
| Campo Mourão (safrinha) | simulado | simulado | simulado | simulado | simulado | X |
| Cristalina | X | X | X | simulado | simulado | X |
| Dourados (safrinha) | X | X | X | Simulado | X | X |
| Goianésia | X | X | X | Simulado | simulado | X |
| Goiânia | X | X | X | X | X | simulado |
| Goiânia (safrinha) | X | X | X | Simulado | X | X |
| Ipameri | X | X | X | X | X | X |
| Itumbiara | X | X | X | Simulado | simulado | X |
| Londrina | X | X | X | simulado | simulado | X |
| Londrina (safrinha) | X | X | X | simulado | simulado | X |
| Maracaju | X | X | simulado | simulado | simulado | X |
| Maracaju (safrinha) | X | X | X | simulado | simulado | X |

**Tabela 3.8**
**Relação de variáveis sobre ensaios.**

| Cidade | AP | AE | Ac% | ED% | PE | PG |
|---|---|---|---|---|---|---|
| Montividiu (safrinha) | X | X | simulado | X | X | X |
| Morrinhos | X | X | X | X | X | simulado |
| Palmeiras de Goiás | simulado | simulado | simulado | X | X | simulado |
| Palotina | X | X | X | simulado | simulado | X |
| Palotina (safrinha) | X | X | X | simulado | simulado | X |
| Patos de Minas | X | X | X | X | X | X |
| Piracicaba | X | X | X | X | X | X |
| Planaltina | X | X | X | X | X | X |
| Ponta Porã | X | X | X | simulado | X | X |
| Ponta Porã (safrinha) | X | X | X | simulado | X | X |
| Porangatu | X | X | X | X | X | X |
| Rio Verde | X | X | X | X | X | simulado |
| Rio Verde (safrinha) | X | X | X | simulado | X | X |
| São Raimundo das Mangabeiras | X | X | X | X | X | X |
| Sete Lagoas | X | X | X | X | X | simulado |
| Sete Lagoas, corrigido | X | X | X | X | X | simulado |
| Sooretama | X | X | X | X | X | simulado |
| Uberlândia | X | X | X | X | X | simulado |

### 3.3.2 Dados de clima

Os dados de clima foram coletados em instituições públicas e através de solicitações diretas a instituições de pesquisa, conforme descrito na Tabela 3.9. Foram coletados dados diários, a partir da data de plantio até a data de colheita. Em determinados casos, por ausência da data de colheita, foram coletados dados até 180 dias após o plantio, pois, essa quantidade de dias é superior ao ciclo da cultura do milho.

**Tabela 3.9**
**Fontes dos dados de clima**

| Cidade | Fonte | Endereço Web |
|---|---|---|
| Assis, Birigui, Baixo Grande Ribeiro, Cristalina, Dourados, Goianésia, Ipameri, Montividiu, Morrinhos, Palmeiras de Goiás, Piracicaba, Planaltina, São Raimundo das Mangabeiras, Sooretama, Uberlândia | Sistema de Monitoramento Agrometeorológico – Agritempo | www.agritempo.gov.br |
| Brasilândia do Sul, Campo Mourão, Londrina, Palotina | Instituto Agronômico do Paraná - IAPAR | www.iapar.br |

**Tabela 3.9**
**Fontes dos dados de clima**

| Cidade | Fonte | Endereço Web |
|---|---|---|
| Maracaju, Ponta Porã | Instituto Nacional de Meteorologia - INMET | www.inmet.gov.br |
| Goiânia, Itumbiara, , Patos de Minas, Porangatu, Rio Verde | Instituto Nacional de Pesquisas Espaciais - INPE | www.cptec.inpe.br |
| Sete Lagoas | Embrapa Milho e Sorgo | www.cnpms.embrapa.br |

As cidades listadas na Tabela 3.10 não possuem estações meteorológicas. Segundo especialistas em meteorologia consultados, a área de cobertura de estação meteorológica é de 25 km de raio, portanto, se na cidade não houver estação, pode-se considerar os dados de uma estação dentro deste raio como dados locais.

Através da utilização do software MapInfo[5] foram selecionadas as estações mais próximas dessas cidades. Em muitos casos a distância da estação mais próxima foi superior a 25 km, porém, como não há outra fonte desta informação, optou-se por utilizar estes dados.

**Tabela 3.10**
**Relação de cidades sem estação meteorológica**

| Cidade do ensaio | Estação Mais Próxima | Distância Geodésica (km) |
|---|---|---|
| Assis | Cândido Mota | 10 |
| Baixo Grande Ribeiro | Bom Jesus | 165 |
| Brasilândia do Sul | Umuarana | 52 |
| Birigui | Votuporanga | 103 |
| Campo Mourão | Nova Cantu | 72 |
| Dourados | Ponta Porã | 101 |
| Goianésia | Pirenópolis | 62 |
| Maracaju | Ponta Porã | 117 |
| Montividiu | Rio Verde | 47 |
| Palmeiras de Goiás | Varjão | 41 |
| Planaltina | Formosa | 31 |
| São Raimundo das Mangabeiras | Balsas | 83 |
| Sooretama | Linhares | 22 |
| Uberlândia | Uberaba | 99 |

Dados sobre radiação solar e vento, foram encontrados em apenas quatro estações e dados sobre umidade relativa do ar em 10 estações. Foi observado que dados sobre

[5] MapInfo Desktop Mapping Software for Windows versão 5.0

temperatura mínima, temperatura máxima e precipitação estão disponíveis em todas as estações meteorológicas.

Através de entrevistas com especialistas, foi ponderada a não utilização dos dados de radiação solar, vento e umidade relativa do ar. Foi estabelecido que a ausência de qualquer variável imporia restrições aos resultados obtidos, mas informações sobre vento e radiação solar seriam menos impactantes do que a umidade relativa do ar, de acordo com especialistas em fisiologia vegetal consultados. Sendo assim, a utilização de dados sobre radiação solar e vento foi descartada, porém a ausência de informações sobre umidade relativa do ar poderia comprometer a confiabilidade e validade dos resultados obtidos na etapa de *clusterização*, objetivo maior deste trabalho.

Em face dessa constatação, ficou claro que os dados sobre umidade relativa do ar não poderiam ser simplesmente excluídos do modelo e que maiores esforços deveriam ser envidados na obtenção dos mesmos. Assim sendo, buscou-se obter estes dados em fontes provenientes de instituições privadas e governamentais, as quais informaram não dispor dos mesmos.

Diante da ausência das informações sobre umidade relativa do ar, surgiu a necessidade de estimar estes dados através da utilização de um modelo computacional neural. O processo utilizado para geração dos dados de umidade relativa do ar é descrito de forma pormenorizada no Capítulo 4.

### 3.3.3 Preparação dos dados de clima

Ao final da etapa de levantamento de dados climáticos, obteve-se os dados diários de temperatura mínima, temperatura máxima e pluviosidade para todas as cidades. Os dados de umidade relativa do ar foram coletados diariamente, somente para as cidades de Brasilândia do Sul, Campo Mourão, Goiânia, Itumbiara, Londrina, Palotina, Patos de Minas, Porangatu, Rio Verde e Sete Lagoas. Para as demais cidades, foi gerada a média mensal da umidade relativa, conforme descrito no capítulo 4.

Conforme dito anteriormente, os dados climáticos seriam utilizados de forma agrupada, obedecendo às etapas ou estádios mais importantes do desenvolvimento da planta, de acordo com a Tabela 3.6. Foi definido juntamente com os especialistas em agrometeorologia e fisiologia vegetal, que se utilizaria o valor médio em cada estádio, para os

dados de temperatura mínima, temperatura máxima e umidade relativa. A pluviosidade seria utilizada de forma acumulada, para cada estádio. A Tabela 3.11 mostra como foram agrupados os dados climáticos.

**Tabela 3.11**
**Critério de agrupamento de dados climáticos**

| Estádio | Tempo médio em que ocorre (em dias após o plantio) | Período utilizado para cálculo (em dias após o plantio) | Forma de cálculo para Tmim, Tmax e UR | Forma de cálculo para pluviosidade |
|---|---|---|---|---|
| VE | 5 | 1 a 5 | Média dos dias 1 a 5 | Soma dos dias 1 a 5 |
| V3 | 12 | 6 a 12 | Média dos dias 6 a 12 | Soma dos dias 6 a 12 |
| V12 | 45 | 13 a 45 | Média dos dias 13 a 45 | Soma dos dias 13 a 45 |
| Florescimento | 55 | 46 a 55 | Média dos dias 46 a 55 | Soma dos dias 46 a 55 |
| Enchimento 1 (R2) | 65 | 56 a 65 | Média dos dias 56 a 65 | Soma dos dias 56 a 65 |
| Enchimento 2 (R4) | 90 | 66 a 90 | Média dos dias 66 a 90 | Soma dos dias 66 a 90 |
| Enchimento 3 (R5) | 100 | 91 a 100 | Média dos dias 91 a 100 | Soma dos dias 91 a 100 |

Planilhas eletrônicas foram utilizadas para o agrupamento dos dados climáticos. Ao final desta etapa, cada cidade tinha o conjunto de valores de Tmin, Tmax, Pluviosidade e UR, para cada um dos sete estádios.

### 3.3.3 Dados de solo

A fonte ideal para se obter os dados de solo é a análise de solo dos locais onde ocorreram os plantios. Foi verificado junto aos responsáveis pelos ensaios sobre a existência destas análises, porém, em nenhum dos locais havia análise de solo disponível. Considerando a indisponibilidade destas análises, a alternativa que se encontrou foi buscar estes dados em levantamentos e mapas de solos.

Os levantamentos de solos utilizados foram elaborados pela Embrapa e são compostos de mapas e livros onde são detalhadas todas as características dos solos estudados. A partir dos mapas têm-se os tipos de solos de cada cidade. Através de tabelas descritivas de cada tipo de solo obtém-se a textura e a saturação por bases, além de informações necessárias para se calcular a água disponível.

A textura é dada a partir da quantidade de argila presente, conforme tabela abaixo:

**Tabela 3.12**
**Classificação da Textura do solo**

| Teor de Argila | Textura |
|---|---|
| = 15 | arenosa |
| > 15 = 35 | média |
| > 35 = 60 | argilosa |
| > 60 | muito argilosa |

Através dos mapas de solos pode-se facilmente constatar a diversidade de tipos de solos existentes em cada cidade. Considerando que não havia informação precisa sobre o tipo de solo onde foram realizados os ensaios, optou-se por utilizar o tipo predominante na região.

É importante ressaltar que, pela escala em que são trabalhadas, as informações constantes nos levantamentos de solos não possuem a precisão e a confiabilidade das informações de análises de solos feitas nos locais onde se realizaram o plantio.

### 3.3.3.1 Preparação dos dados de solos

Considerando que o valor da textura é dado em função da quantidade de argila presente, conforme classificação descrita na Tabela 3.12, optou-se por utilizar então o próprio valor da argila, uma vez que são valores seqüenciais e assim a distância ou similaridade entre um solo e outro ficaria melhor quantificada.

O valor da água disponível pode ser obtido, de acordo com Assad *et al*. (2001), através da equação abaixo:

$$AD = a + b \times (AT)^3$$

onde:

*AD* = água disponível

*a = 12,76278562* (parâmetro de ajuste)

*b = -9,87626e-06* (parâmetro de ajuste)

*AT = areia total*

A altitude do local de plantio foi obtida nas planilhas descritivas de cada ensaio.

A Tabela 3.13 mostra, ao final da etapa de tratamento dos dados, os tipos de solos e suas características.

**Tabela 3.13**
**Informações sobre os solos predominantes nas cidades onde se realizaram os plantios**

**(Continua)**

| Cidade | Altitude | Solo | Descrição | Textura | Saturação por Bases (V%) | Água Disponível |
|---|---|---|---|---|---|---|
| Assis | 546 | Lvd1 | Latossolo vermelho-escuro distrofico | Arenosa (argila = 9) | 26 | 6,483 |
| Baixo Grande Ribeiro | 325 | Lld1 | Latossolo vermelho-escuro distrofico | Argilosa (argila = 56) | 6 | 12,496 |
| Birigui | 406 | Lvd7 | Latossolo vermelho-escuro distrofico | Arenosa (argila = 9) | 26 | 6,483 |
| Brasilândia do Sul | 378 | Pe3 | Podzólico vermelho-amarelo eutrófico Tb | Arenosa (argila = 9) | 91 | 8,429 |
| Campo Mourão | 585 | Lrd5 | Latossolo roxo distrófico | Muito Argiloso (argila = 73) | 72 | 12,761 |
| Cristalina | 1189 | Lld12 | Latossolo vermelho-escuro distrofico plano e suave ondulado. | Argilosa (argila = 59) | 8 | 12,695 |
| Dourados | 430 | Lrd1 | Latossolo roxo distrófico | Muito argilosa (argila = 61) | 77 | 12,705 |
| Goianésia | 640 | Lvd6 | Latossolo vermelho-escuro distrofico + Latossolo roxo distrófico e eutrófico | Argilosa (argila = 45) | 24 | 12,031 |
| Goiânia | 749 | Lvd7 | Latossolo vermelho-escuro distrofico + Podzólico vermelho-amarelo distrófico Tb | Argilosa (argila = 45) | 24 | 12,031 |
| Ipameri | 764 | Lld6 | Latossolo vermelho-amarelo distrófico + Latossolo vermelho-escuro distrófico | Arenosa (argila = 12) | 7 | 7,118 |
| Itumbiara | 448 | Lrde1 | Latossolo roxo distrófico e eutrófico | Argilosa (argila = 49) | 7 | 12,375 |
| Londrina | 585 | Lre8 | Latossolo roxo eutrófico A | Muito Argiloso (argila = 69) | 79 | 12,746 |
| Maracaju | 384 | Lrd1 | Latossolo roxo distrófico | Muito argilosa (argila = 61) | 77 | 12,705 |
| Montividiu | 821 | Lvd4 | Latossolo vermelho-escuro distrofico + Latossolo vermelho-amarelo distrófico + areias quartzosas distróficas | Argilosa (argila = 42) | 13 | 11,978 |
| Morrinhos | 771 | Lvd2 | Latossolo vermelho-escuro distrofico + Latossolo vermelho-amarelo distrófico | Argilosa (argila = 42) | 13 | 11,978 |
| Palmeiras de Goiás | 596 | Lvd7 | Latossolo vermelho-escuro distrofico + Podzólico vermelho-amarelo distrófico Tb | Argilosa (argila = 45) | 24 | 12,031 |
| Palotina | 333 | LRe1 | Latossolo roxo eutrófico A | Muito Argiloso (argila = 80) | 48 | 12,762 |
| Patos de Minas | 832 | Cd5 / Lld7 | Cambissolo distrófico Álico Fraco | Argilosa (agila = 51) | 8 | 12,695 |
| Piracicaba | 596 | Pd10 | Podzólico vermelho-amarelo distrófico | Média (argila = 28) | 23 | 10,522 |
| Planaltina | 944 | Cd4 | Cambissolo distrófico Álico Fraco | Argilosa (argila = 42) | 12 | 12,496 |

**Tabela 3.13**
**Informações sobre os solos predominantes nas cidades onde se realizaram os plantio**

**(Conclusão)**

| Cidade | Altitude | Solo | Descrição | Textura | Saturação por Bases (V%) | Água Disponível |
|---|---|---|---|---|---|---|
| Ponta Porã | 665 | Lrd1 | Latossolo roxo distrófico | Muito argilosa (argila = 61) | 77 | 12,705 |
| Porangatu | 396 | Lld7 | Latossolo vermelho-amarelo distrófico + Podzólico vermelho-amarelo distrófico Tb | Arenosa (argila = 6) | 67 | 6,035 |
| Rio Verde | 715 | Lvd4 | Latossolo vermelho-escuro distrofico + Latossolo vermelho-amarelo distrófico + areias quartzosas distróficas | Argilosa (argila = 42) | 13 | 11,978 |
| São Raimundo das Mangabeiras | 225 | | Latossolo amarelo álico | Argilosa (argila = 47) | 29 | 12,684 |
| Sete Lagoas | 761 | Lvd2 | Latossolo vermelho-Amarelo Distrófico, húmico epieutrófico, textura muito argilosa fase floresta tropical subperenifólia relevo suave ondulado | Muito Argilosa (argila = 63) | 56 | 12,671 |
| Sooretama | 59 | La1 | Argissolo amarelo Distrófico típico A moderado textura arenosa/ | Média (argila = 29) | 19 | 7,319 |
| Uberlândia | 863 | Lvd1 | Latossolo Vermelho-escuro distrófico | Argilosa (argila = 41) | 29 | 12,082 |

## 3.4 Montagem da base de dados

De posse dos dados, o próximo passo foi montar a base de dados no formato apropriado para a etapa de mineração de dados.

A estrutura do arquivo requerido pelo algoritmo minerador é de tal forma que, em cada linha se tenha todos os dados dos elementos a se agrupar. Desta forma, os dados de ensaios, clima e solos de cada experimento devem estar na mesma linha. Essa inversão de colunas por linhas é conhecida como pivotagem reversa. A estrutura do arquivo é mostrada na Figura 3.3.

| | **Dados da Cultivar 1** | | | | | | **Dados da Cultivar 2** | |
|---|---|---|---|---|---|---|---|---|
| Cidade | Altura da planta (cm) | Altura da espiga (cm) | Acamadas + quebradas (%) | Espigas doentes (%) | Peso de espigas (kg/ha) | Peso de grãos (kg/ha) | Altura da planta (cm) | … |

| … | **Dados da Cultivar 32** | | | | | | … |
|---|---|---|---|---|---|---|---|
| … | Altura da Planta (cm) | Altura da Espiga (cm) | Acamadas + quebradas (%) | Espigas doentes (%) | Peso de espigas (kg/ha) | Peso de grãos (kg/ha) | … |

| … | **Dados do solo** | | | **Dados de Clima - Estádio VE** | | | | | … |
|---|---|---|---|---|---|---|---|---|---|
| … | Altitude | Saturação por bases (V%) | Textura | Água disponível | Temperatura mínima | Temperatura máxima | Precipi-tação | Umida-de rela-tiva | … |

| … | **Dados de Clima – Estádio Enchimento 3** | | | |
|---|---|---|---|---|
| … | Temperatura mínima | Temperatura máxima | Precipitação | Umidade relativa |

Figura 3.3 – Estrutura do arquivo utilizado na etapa de mineração de dados

Considerando que existem 32 cultivares, 6 atributos por cultivar, 4 atributos de solo, 7 fases (estádios) de amostragem de dados climáticos com 4 atributos por fase e 35 ensaios, têm-se:

(32 x 6) + 4 + (7 x 4) = 224 atributos por cidade

Considerando o número de ensaios, o arquivo terá 35 registros (tuplas) com 224 atributos em cada registro.

A fim de facilitar o manuseio dos dados, foi utilizado o software Microsoft Excel 2007 para montar o banco de dados de acordo com estrutura acima.

Para utilização do algoritmo minerador, é necessário normalizar esses dados, conforme já citado. Para essa tarefa a equação 2.1 apresentada na seção 2.2.2.3 foi utilizada.

# 4. REPRESENTAÇÃO NEURAL DE DADOS CLIMÁTICOS – MODELO PARA ESTIMAR A UMIDADE RELATIVA DO AR

## 4.1 Introdução

No capítulo anterior foi descrito o modelo da interação Planta - Solo - Clima. Foi levantado um conjunto de características (atributos) sobre cada um desses componentes que, na visão de especialistas do domínio (fisiologia vegetal, melhoramento genético e meteorologistas), mais influenciam o desenvolvimento da planta de milho.

Sobre os dados referentes ao clima, foi verificado, em consultas a especialistas, que a radiação solar somente é medida em estações meteorológicas mais modernas, o que não é o caso da maioria das cidades consideradas neste trabalho. Foi constatado ainda que os dados de vento e umidade relativa do ar são inconsistentes ou estão ausentes em 20 das 30 estações utilizadas. Foi observado que dados sobre temperatura mínima, temperatura máxima e precipitação estão disponíveis em todas as estações.

Diante da ausência de dados sobre radiação solar e vento optou-se por não utilizar esses dados. Sobre os dados de umidade relativa do ar, constatou-se junto aos especialistas que a não utilização poderia comprometer a confiabilidade dos resultados obtidos na etapa de mineração de dados.

Através de contatos com instituições privadas e governamentais, tentou-se obter os dados de umidade relativa do ar, para as 20 cidades que não dispunham dessas informações, sem, no entanto, obter sucesso. Como resultado desses contatos, obteve-se uma coleção de dados históricos de 255 estações meteorológicas, ao longo de 12 anos. Estes dados contêm a leitura diária dos dados coletados por estas estações, existindo entretanto, lacunas com datas não coletadas. Medidas de umidade relativa do ar, além de outros dados foram registradas de 1995 a 2006. A relação destas estações com a quantidade de dias de leitura de cada ano é listada no Apêndice A.

Das 30 cidades onde os ensaios foram realizados, apenas oito cidades constavam nesta coleção de dados. Diante da ausência das informações de umidade relativa do ar, surgiu a necessidade de estimar estes dados através da utilização de um modelo matemático.

A utilização de redes neurais artificiais (RNA) para representar processos físicos é amplamente relatada na literatura. Neste trabalho RNA foram utilizadas para estimar os dados de umidade relativa do ar.

Com base nos dados observados ao longo dos 12 anos, das 255 estações, uma representação neural foi proposta para relacionar os principais dados disponíveis com a umidade relativa do ar.

## 4.2 A representação neural para estimar dados de umidade relativa do ar

O objetivo da utilização de RNA neste trabalho é estimar os dados de umidade relativa do ar para as cidades onde foram realizados os ensaios e não se dispunham destes dados.

No processo de treinamento da rede foram fornecidos dados de entrada e a saída desejada, no caso, a umidade relativa do ar. Depois de concluído o treinamento e o processo de validação, a rede, conhecendo a relação entre entrada e saída e valendo-se de sua capacidade de generalização, foi utilizada para estimar os dados de umidade das cidades desejadas.

No processo de aprendizado supervisionado, um conjunto de pares de entrada/saída é fornecido à rede a fim de que a mesma consiga aprender a relação entre eles.

A Tabela 4.1 mostra um conjunto de dados de treinamento com $N$ exemplos e $M$ atributos. As colunas ($A_1 \ldots A_M$) representam os atributos ou variáveis de entrada e a coluna Y ($Y_1 \ldots Y_N$) representa a saída. As linhas ($E_1 \ldots E_N$) representam os exemplos.

Na etapa de treinamento, a rede é alimentada com um conjunto de exemplos $E = \{E_1, E_2, \ldots E_N\}$, no qual cada exemplo $E_i \in E$ é uma tupla $(\vec{x}_i, y_i)$ sendo $\vec{x}_i$ um vetor de valores de atributos (entradas) do exemplo $E_i$ e $Y_i$ um valor de saída (resultado) associado ao exemplo $E_i$.

**Tabela 4.1**
**Conjunto de treinamento de uma RNA**

| | $A_1$ | $A_2$ | ... | $A_M$ | Y |
|---|---|---|---|---|---|
| $E_1$ | $X_{11}$ | $X_{12}$ | ... | $X_{1M}$ | $Y_1$ |
| $E_2$ | $X_{21}$ | $X_{22}$ | ... | $X_{2M}$ | $Y_2$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $E_N$ | $X_{N1}$ | $X_{N2}$ | ... | $X_{NM}$ | $Y_N$ |

O conjunto de dados utilizado para o treinamento da rede é referente às 255 estações meteorológicas para os 12 anos observados.

Os dados disponíveis das cidades que deseja-se estimar a umidade são: latitude (lat), longitude (long), altitude (alt), temperatura máxima (Tmax), temperatura mínima (Tmin). Assim estes dados foram utilizados como entrada no processo de treinamento.

No conjunto de dados das 255 estações foi constatada a existência de outros parâmetros (temperatura às 12:00/18:00/00:00, temperatura de bulbo seco às 12:00/18:00/00:00, velocidade do vento, pressão atmosférica). A utilização destes dados poderia contribuir positivamente na representação neural, porém, sua utilização foi descartada porque estes dados não estão disponíveis para as localidades que se desejava estimar a umidade.

A umidade relativa do ar neste caso pode ser descrita como uma função não linear representada abaixo:

$$F(lat, long, alt, Tmin, Tmax) \xrightarrow{RNA} UR$$

Sendo:

*lat* = latitude

*long* = longitude

*alt* = altitude

*Tmin* = temperatura mínima

*Tmax* = temperatura máxima

*UR* = umidade relativa do ar

A RNA utilizada tem, portanto, cinco entradas e um neurônio na camada de saída e pode ser esquematizada conforme Figura 4.1.

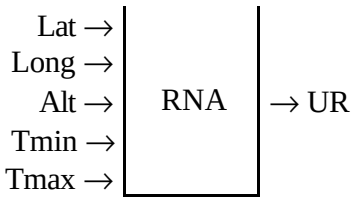


Figura 4.1 – RNA para estimar umidade

Neste trabalho, a rede neural utilizada é do tipo multicamadas, sendo uma camada correspondente às entradas da rede, uma camada de saída com um neurônio e entre elas uma camada intermediária conhecida como camada oculta. A quantidade de neurônios na camada oculta segue critérios empíricos. Segundo Braga *et al.* (2003), este número depende fortemente da distribuição dos padrões de treinamento e da validação da rede. Vários métodos foram propostos para sua definição, sendo os mais utilizados descritos a seguir:

1. Definir o número de unidades em função do número de entradas e saídas;
2. Utilizar um número de conexões dez vezes menor que o número de exemplos.

Neste trabalho o primeiro método foi utilizado, mais especificamente um critério conhecido como Teorema de Kolmogorov (Hecht-Nielsen, 1989) (Kovacs, 1996), que diz que uma RNA Perceptron multicamadas (MLP) com uma camada oculta contendo "2n + 1" neurônios pode representar qualquer função, onde "n" é o número de entradas. Neste caso n = 5, portanto, a camada oculta conterá 11 neurônios.

### 4.2.1 Preparação dos dados do conjunto de treinamento

Os dados históricos obtidos compreendem o período de 01/01/1995 a 18/06/2006 e estavam armazenados em meio digital. Para cada dia existia um arquivo separado, perfazendo um total de 4186 arquivos. Cada arquivo continha dados das 255 estações.

A localização das estações meteorológicas em todo o país pode ser observada na Tabela 4.2 e na Figura 4.2.

**Tabela 4.2**
**Distribuição das estações meteorológicas por UF**

| UF | # estações | UF | # estações | UF | # estações | UF | # estações | UF | # estações |
|---|---|---|---|---|---|---|---|---|---|
| AC | 4 | DF | 2 | MT | 12 | RJ | 11 | SE | 4 |
| AL | 3 | ES | 4 | PA | 18 | RN | 7 | SP | 19 |
| AM | 16 | GO | 14 | PB | 5 | RO | 3 | TO | 1 |
| AP | 2 | MA | 10 | PE | 8 | RR | 4 | | |
| BA | 1 | MG | 40 | PI | 11 | RS | 21 | | |
| CE | 11 | MS | 5 | PR | 10 | SC | 9 | | |

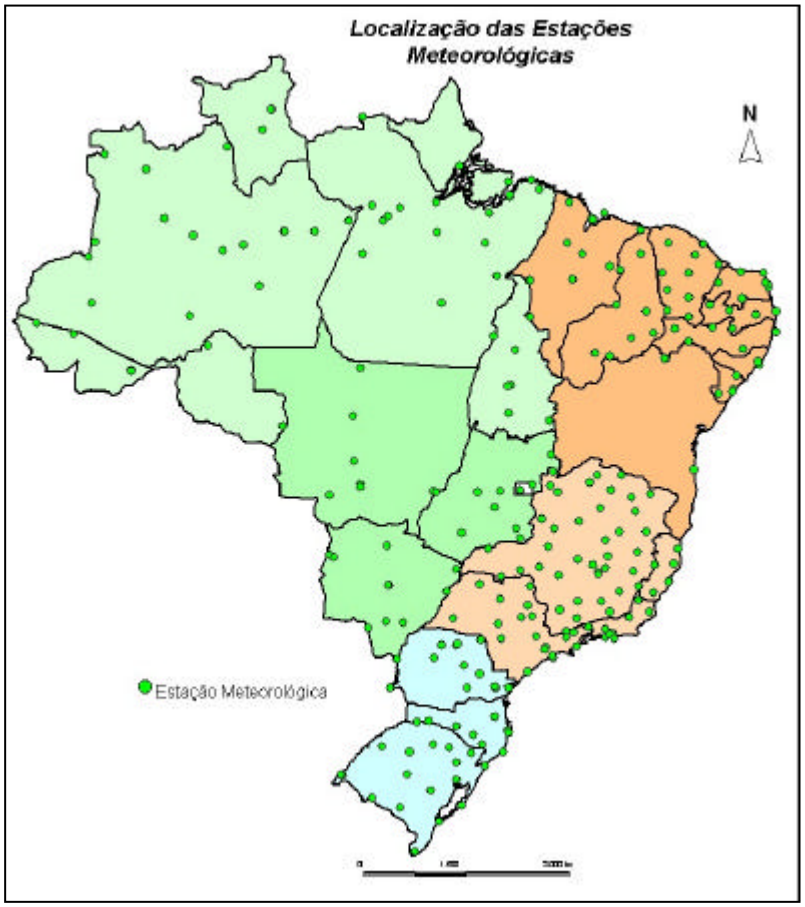


Figura 4.2 – Localização das estações meteorológicas utilizadas

Os procedimentos descritos abaixo foram executados para selecionar os dados utilizados na etapa de treinamento da rede neural. No Apêndice B estes procedimentos estão descritos de forma detalhada.

**Passo 1 - Separação dos arquivos por estações e anos**

Nesta etapa as leituras diárias foram agrupadas por estação e por ano, obtendo-se ao final um conjunto de arquivos onde cada arquivo contêm todas as leituras de uma estação, em um ano.

**Passo 2 - Remoção de registros com inconsistências/suspeitos**

Nesta etapa foi feita uma profunda checagem nos dados. Foi feita também uma análise de *outliers* e registros inconsistentes ou com dados faltantes/suspeitos foram removidos.

**Passo 3 - Redução da quantidade de amostras**

Após a remoção de registros inconsistentes e/ou suspeitos, foi constatado um número de amostras elevado (648104 registros), o que poderia inviabilizar o treinamento da RNA pelo custo computacional requerido. Nesta etapa foram adotados procedimentos para reduzir a quantidade de amostras mantendo a representatividade do conjunto. Para isto foram utilizadas técnicas de análise estatística para determinar a quantidade ideal de amostras.

**Passo 4 – Seleção de amostras representativas e montagem dos conjuntos de treinamento**

Nesta etapa foram selecionadas as estações com quantidade de dados estatisticamente representativos. Estas estações foram agrupadas em 12 conjuntos contendo, cada um, dados referentes a um mês.

Ao final dos procedimentos acima, obteve-se o conjunto de treinamento formado por 12 conjuntos de dados. Cada conjunto contém dados médios mensais de estações meteorológicas de todo o Brasil, conforme Tabela 4.3.

**Tabela 4.3**
**Quantidade de estações válidas por mês/UF do conjunto de treinamento**

| Mês | AC | AL | AM | AP | CE | DF | ES | GO | MA | MG | MS | MT | PA | PB | PE | PI | PR | RJ | RN | RO | RR | RS | SC | SE | SP | TO | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Jan** | 3 | 1 | 14 | 2 | 8 | 1 | 2 | 6 | 8 | 17 | 4 | 4 | 15 | 2 | 4 | 6 | 7 | 3 | 3 | 1 | 3 | 10 | 2 | 1 | 6 | 4 | **137** |
| **Fev** | 3 | 1 | 14 | 1 | 9 | 1 | 2 | 6 | 9 | 17 | 4 | 4 | 15 | 3 | 5 | 6 | 7 | 4 | 3 | 1 | 3 | 10 | 2 | 1 | 6 | 4 | **141** |
| **Mar** | 3 | 1 | 14 | 1 | 10 | 1 | 3 | 7 | 9 | 17 | 5 | 4 | 16 | 4 | 6 | 9 | 7 | 4 | 3 | 1 | 3 | 10 | 2 | 3 | 6 | 4 | **153** |
| **Abr** | 3 | 1 | 14 | 1 | 10 | 1 | 3 | 6 | 9 | 17 | 5 | 4 | 15 | 4 | 6 | 8 | 7 | 4 | 3 | 1 | 3 | 10 | 2 | 1 | 6 | 4 | **148** |
| **Mai** | 3 | 1 | 14 | 1 | 10 | 1 | 3 | 6 | 9 | 17 | 5 | 4 | 16 | 4 | 5 | 8 | 7 | 4 | 3 | 1 | 3 | 10 | 2 | 1 | 6 | 4 | **148** |
| **Jun** | 3 | 1 | 14 | 2 | 9 | 1 | 3 | 6 | 9 | 17 | 3 | 4 | 16 | 4 | 5 | 7 | 7 | 4 | 3 | 2 | 3 | 10 | 2 | 3 | 6 | 4 | **148** |
| **Jul** | 3 | 1 | 14 | 2 | 10 | 1 | 3 | 6 | 9 | 17 | 2 | 4 | 16 | 4 | 6 | 8 | 7 | 3 | 3 | 1 | 3 | 10 | 2 | 3 | 5 | 4 | **147** |
| **Ago** | 3 | 1 | 14 | 2 | 11 | 2 | 3 | 7 | 9 | 17 | 2 | 5 | 16 | 4 | 6 | 9 | 7 | 3 | 5 | 1 | 3 | 10 | 2 | 3 | 6 | 4 | **155** |
| **Set** | 3 | 1 | 14 | 2 | 11 | 2 | 3 | 7 | 9 | 17 | 0 | 4 | 16 | 4 | 6 | 9 | 5 | 3 | 5 | 1 | 3 | 10 | 2 | 3 | 5 | 4 | **149** |
| **Out** | 3 | 1 | 14 | 2 | 10 | 2 | 3 | 7 | 9 | 17 | 4 | 5 | 16 | 4 | 6 | 8 | 7 | 3 | 5 | 1 | 3 | 10 | 2 | 3 | 5 | 4 | **154** |
| **Nov** | 3 | 1 | 14 | 2 | 10 | 1 | 3 | 7 | 9 | 17 | 3 | 4 | 16 | 4 | 6 | 8 | 6 | 3 | 4 | 2 | 3 | 10 | 2 | 3 | 6 | 4 | **151** |
| **Dez** | 3 | 1 | 14 | 1 | 8 | 1 | 2 | 6 | 9 | 17 | 4 | 4 | 15 | 3 | 5 | 6 | 6 | 4 | 3 | 1 | 3 | 10 | 2 | 2 | 6 | 4 | **140** |

Conforme pode ser visto na Tabela 4.3, das 255 estações iniciais, no pior caso 118 foram desprezadas (janeiro), e no melhor caso, foram desprezadas 100 estações (agosto).

Os procedimentos descritos acima com todos os passos efetuados para se chegar ao conjunto de treinamento encontram-se detalhados no Apêndice B.

### 4.2.2 Treinamento da rede neural

Para o treinamento das RNA foi utilizado o software Matlab versão 6.5. Foram utilizados dois *scripts* de comandos Matlab, sendo um para treinamento da rede, chamado "treina_rna", e outro para operação, chamado "opera_rna". Estes *scripts* encontram-se nos Apêndices C e D.

Considerando os 12 conjuntos de dados, 12 RNA foram treinadas, uma para cada mês. O processo de treinamento requer valores normalizados. Essa operação é feita através da função "*premnmx*", que normaliza todos os valores de entradas e saídas do conjunto de treinamento, em valores entre -1 e 1. A operação inversa é feita através da função "*postmnmx*" do Matlab

Para este trabalho foram comparados os algoritmos de treinamento descritos na Tabela 4.4.

**Tabela 4.4**
**Algoritmos de treinamento utilizados**

| Característica Principal do Algoritmo | Algoritmo de Treinamento (nome MatLab) |
|---|---|
| Atualiza os valores dos pesos e *bias* de acordo com a função gradiente descendente com *momentum* | *Traingdm* |
| Atualiza os valores dos pesos e *bias* de acordo com a função gradiente descendente com *momentum* com taxa de aprendizado adaptativa | *Traingdx* |
| Atualiza os valores dos pesos e *bias* de acordo com a função gradiente descendente e taxa de aprendizado variável | *Traingda* |
| Atualiza os valores dos pesos e *bias* de acordo com a otimização de Levenberg-Marquardt | *Trainlm* |
| Atualiza os valores dos pesos e *bias* de acordo com o algoritmo Rprop (*Resilient Back-propagation*) | *Trainrp* |

Foi utilizado como critério para avaliar o desempenho dos algoritmos de treinamento o "erro máximo" calculado entre os valores reais e os valores simulados pela rede, considerando o mesmo conjunto de dados utilizado no treinamento.
O algoritmo que apresentou o menor erro máximo em todas as 12 redes, foi o algoritmo baseado na otimização de Levenberg-Marquardt (*Trainlm*).

O algoritmo *back-propagation* com função de otimização Levenberg-Marquardt (Levenberg, 1944), (Marquardt, 1963) é considerado um dos algoritmos mais eficientes para treinamento de RNA. Este algoritmo apresenta uma capacidade de convergência alta o que significa que, em poucas épocas de treinamento, os pesos sofrem grandes ajustes. De acordo Barbosa *et al.* (2005), para a aceleração do treinamento esse algoritmo utiliza as derivadas de segunda ordem do erro quadrático em relação aos pesos, ao contrário do algoritmo *back-propagation* tradicional, que considera as derivadas de primeira ordem.

O detalhamento matemático dos diversos algoritmos de treinamento utilizados foge ao escopo deste trabalho, entretanto, informações detalhadas sobre o método Levenberg-Marquardt podem ser obtidas em Queiroz (2005).

O processo de treinamento foi executado durante 400 épocas. Utilizando o algoritmo Levenberg-Marquardt, 400 iterações se mostrou um número suficiente pois a

rede já havia atingido o “mínimo local”, condição essa em que o erro não reduz com o aumento do número de iterações. A Tabela 4.5 mostra os resultados obtidos.

**Tabela 4.5**
**Operação das RNA utilizando dados do treinamento como entrada**

| Mês | Número de Estações | Erro (valor real UR – valor estimado de UR pela RNA | | | |
|---|---|---|---|---|---|
| | | Mínimo | Máximo | Médio | Desvio Padrão |
| Jan | 137 | 0,006 | 3,640 | 0,893 | 0,892 |
| Fev | 141 | 0,004 | 3,508 | 0,918 | 0,872 |
| Mar | 153 | 0,000 | 4,026 | 1,041 | 0,983 |
| Abr | 148 | 0,016 | 4,193 | 1,027 | 0,905 |
| Mai | 148 | 0,001 | 4,766 | 0,988 | 1,150 |
| Jun | 148 | 0,000 | 3,862 | 1,017 | 0,974 |
| Jul | 147 | 0,000 | 4,781 | 1,107 | 0,987 |
| Ago | 155 | 0,011 | 4,798 | 1,223 | 1,078 |
| Set | 149 | 0,001 | 4,480 | 1,163 | 1,097 |
| Out | 154 | 0,012 | 4,864 | 1,171 | 1,013 |
| Nov | 151 | 0,000 | 4,835 | 1,023 | 0,983 |
| Dez | 140 | 0,000 | 3,446 | 0,862 | 0,743 |

Na tabela é possível observar que todos os erros máximos ficaram abaixo de 5%, com um desvio padrão máximo inferior a 1,15%. O maior erro (4,86%) foi observado no mês de outubro, que também registrou o segundo maior erro médio. Os meses de janeiro, fevereiro, maio e dezembro apresentaram erro médio abaixo de 1%. Considerando que o erro em todos os meses foi baixo, não se buscou razões que justificassem o motivo do erro ser maior ou menor em determinado mês.

Os Gráficos 4.1 a 4.12 mostram o comportamento da rede para todos os meses do ano, considerando sempre os mesmos dados utilizados no treinamento.

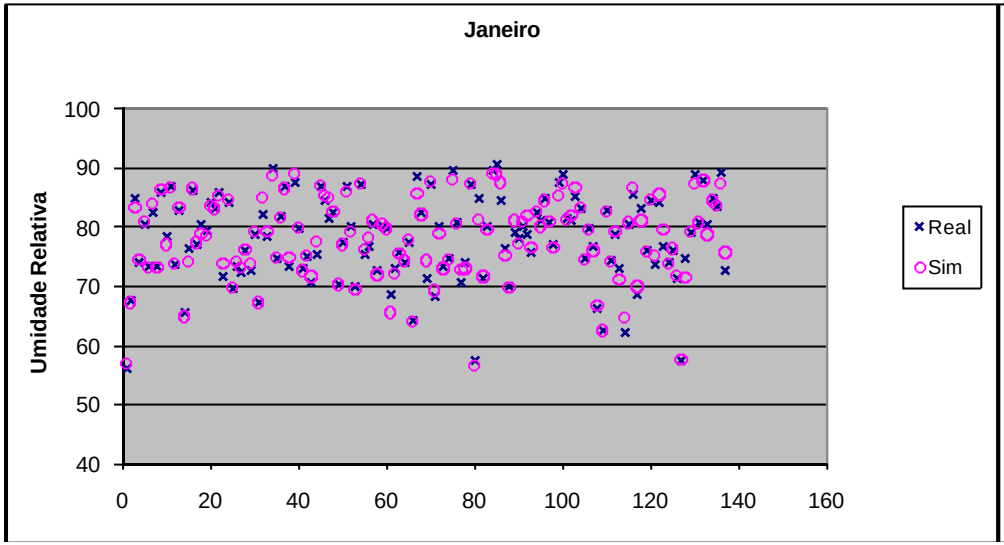


**Gráfico 4.1 – Umidade real X simulada RNA – Janeiro**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 3,640 | 0,8931 | 0,892 |

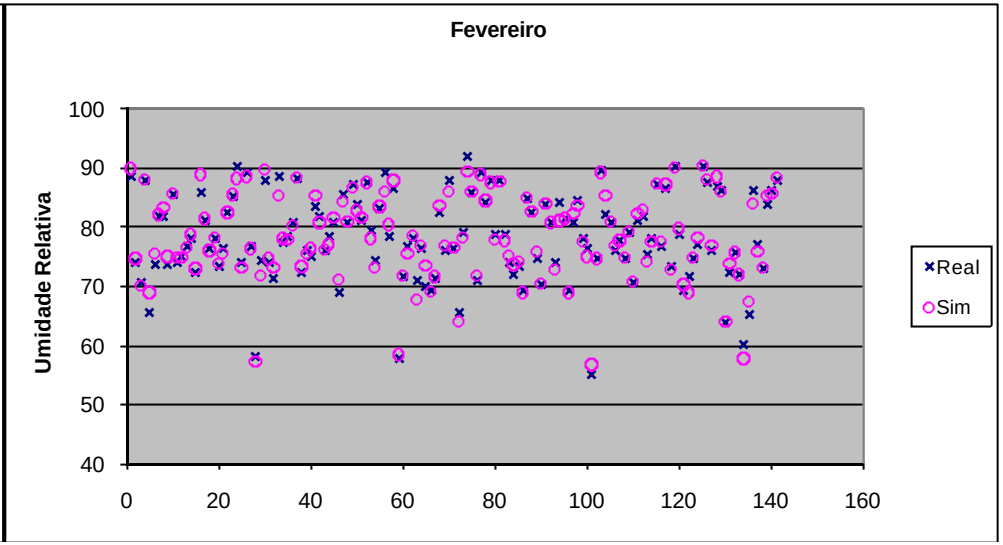


**Gráfico 4.2 – Umidade real X simulada RNA – Fevereiro**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 3,508 | 0,918 | 0,872 |

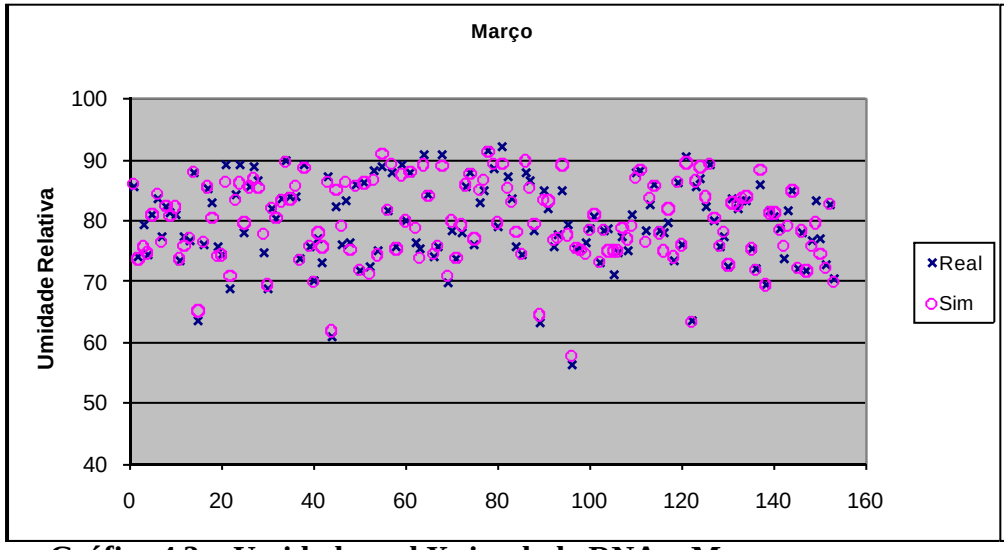


**Gráfico 4.3 – Umidade real X simulada RNA – Março**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 4,026 | 1,041 | 0,983 |

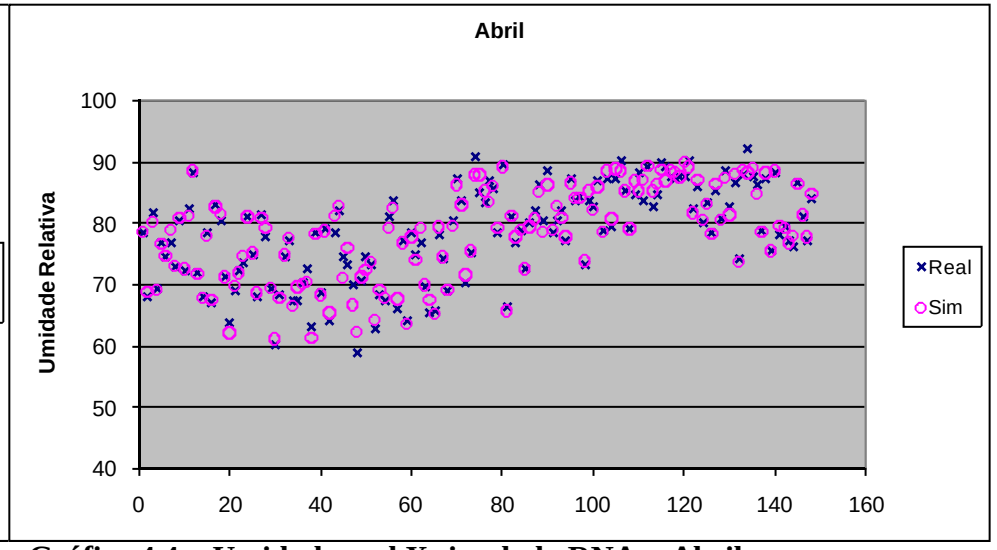


**Gráfico 4.4 – Umidade real X simulada RNA – Abril**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 4,193 | 1,027 | 0,905 |

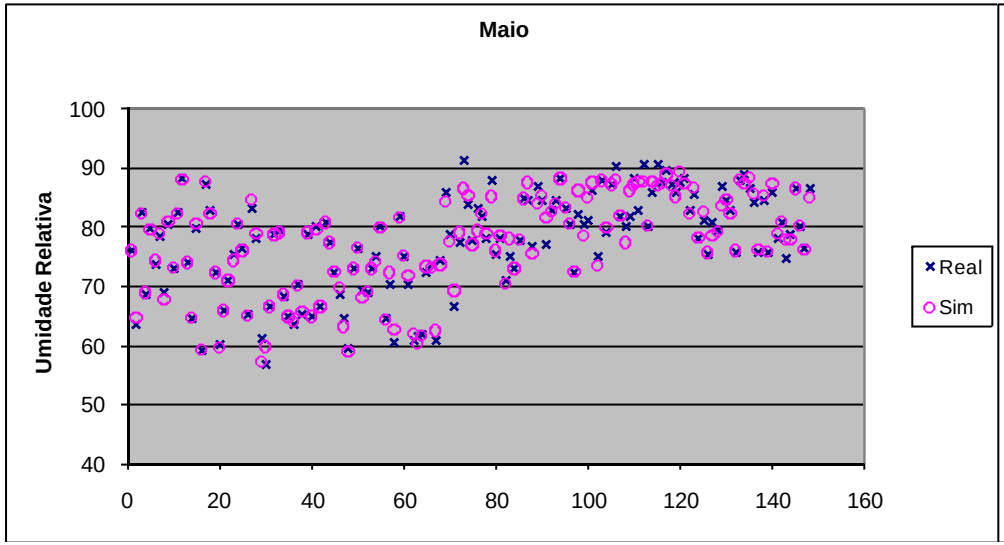


**Gráfico 4.5 – Umidade real X simulada RNA – Maio**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 4,766 | 0,988 | 1,150 |

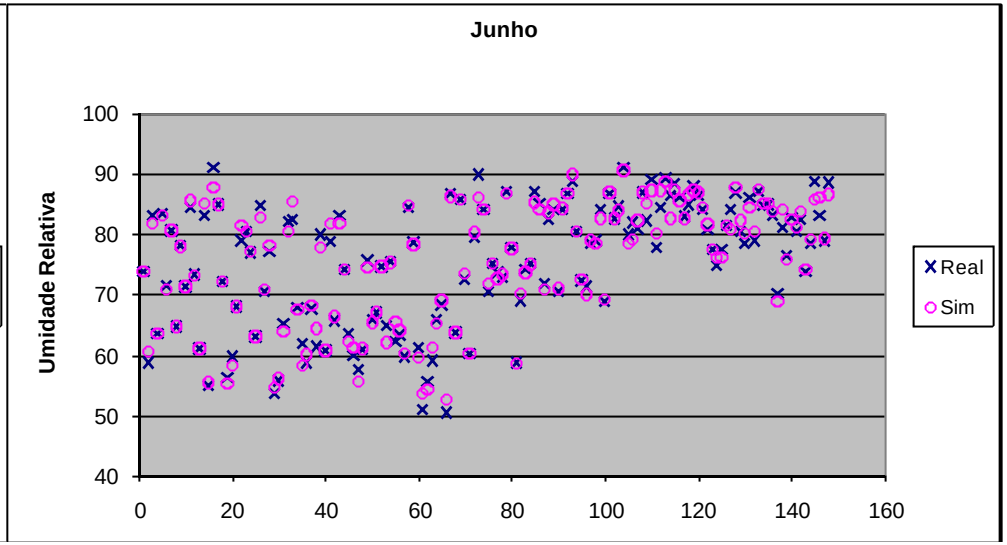


**Gráfico 4.6 – Umidade real X simulada RNA – Junho**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 3,862 | 1,017 | 0,974 |

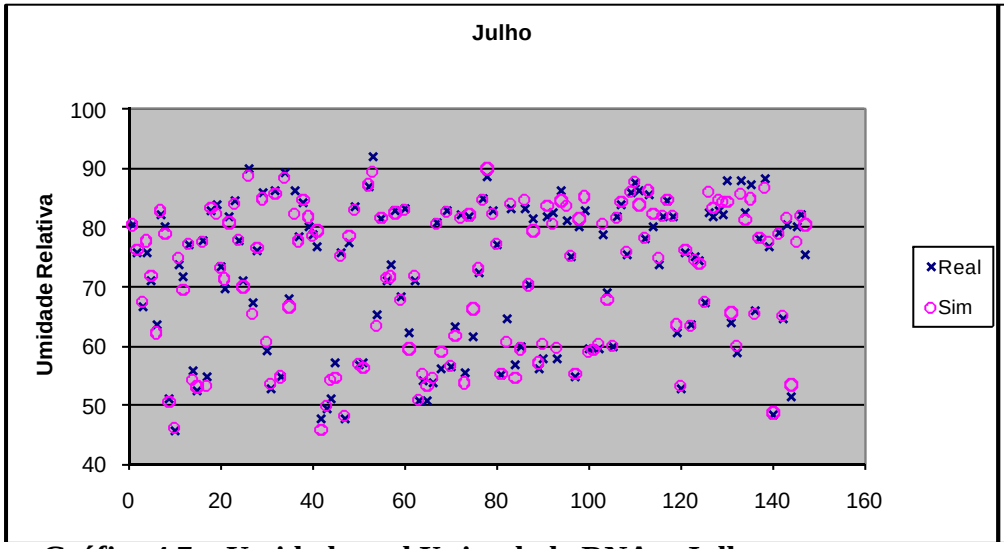


**Gráfico 4.7 – Umidade real X simulada RNA – Julho**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 4,781 | 1,10663 | 0,98673 |

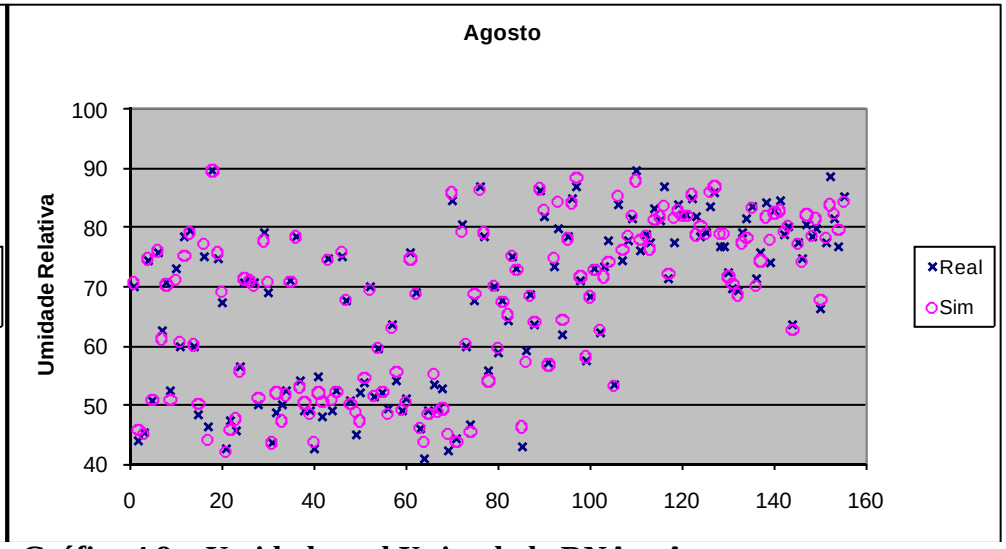


**Gráfico 4.8 – Umidade real X simulada RNA – Agosto**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 4,798 | 1,22295 | 1,07853 |

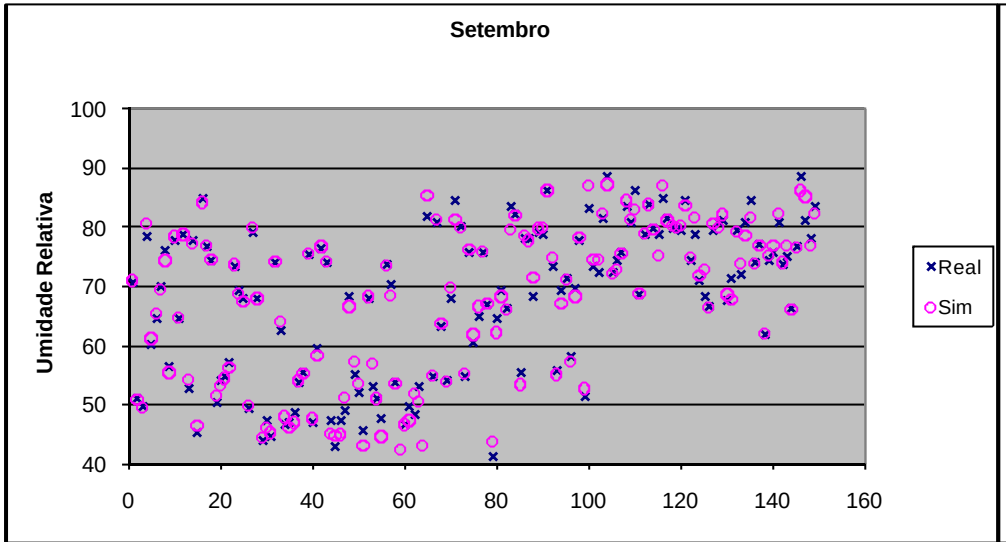


**Gráfico 4.9 – Umidade real X simulada RNA – Setembro**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 4,480 | 1,163 | 1,097 |

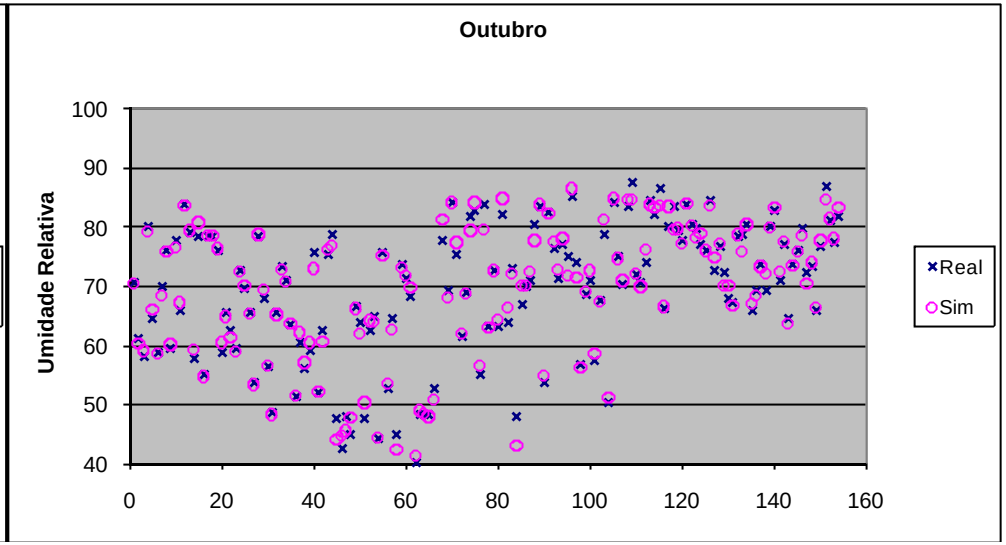


**Gráfico 4.10 – Umidade real X simulada RNA – Outubro**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 4,864 | 1,171 | 1,013 |

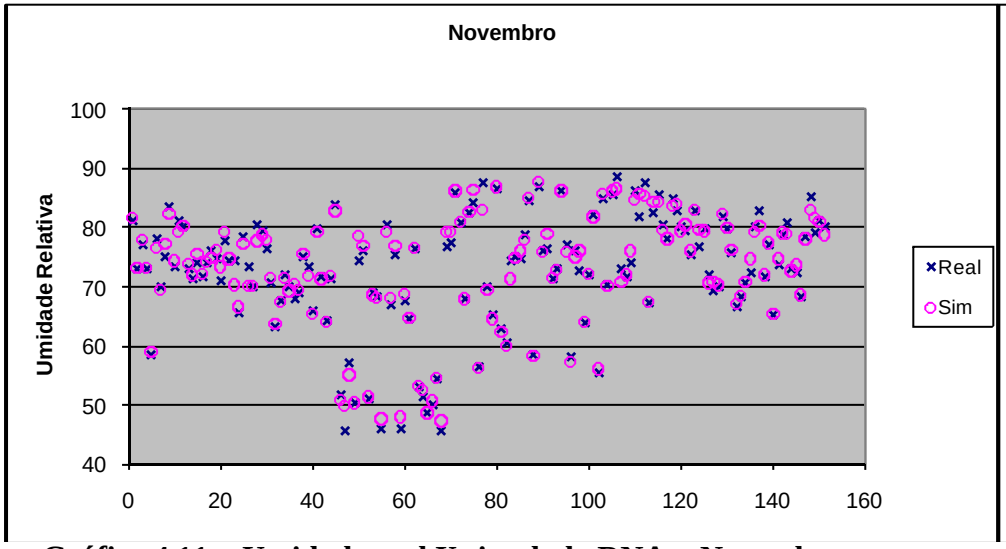


**Gráfico 4.11 – Umidade real X simulada RNA – Novembro**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 4,835 | 1,023 | 0,983 |

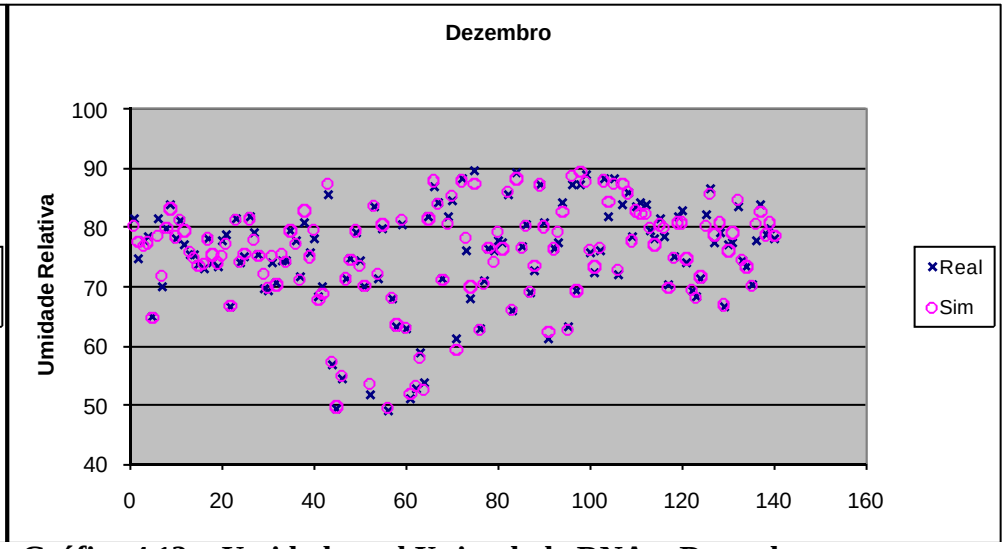


**Gráfico 4.12 – Umidade real X simulada RNA – Dezembro**

| Erro Máximo | Erro Médio | Desvio Padrão |
|---|---|---|
| 3,446 | 0,862 | 0,743 |

### 4.2.3 Validação dos resultados

Para validar as RNA com dados não utilizados no treinamento, foi utilizada a base de dados histórica dos 12 anos de observações, com leituras diárias, com as inconsistências removidas e separados por mês (arquivos janeiro, fevereiro,...,dezembro).

Neste teste, os erros máximos apresentados foram superiores aos obtidos no processo de treinamento, conforme esperado. Os erros médios também foram superiores, mas ainda ficaram dentro de patamares razoáveis, sendo o maior erro médio observado de 12,28%. A Tabela 4.6 mostra o comportamento das RNA utilizando essa base de testes.

**Tabela 4.6**
**Operação das RNA utilizando dados não utilizados no treinamento**

| Mês | Número de Estimativas | Erro (valor real UR – valor estimado de UR pela RNA) | | | |
|---|---|---|---|---|---|
| | | Mínimo | Máximo | Médio | Desvio Padrão |
| Jan | 37013 | 0,001 | 49,564 | 5,854 | 5,009 |
| Fev | 34373 | 0,001 | 48,998 | 8,191 | 6,892 |
| Mar | 37829 | 0,000 | 50,522 | 8,543 | 7,307 |
| Abr | 31544 | 0,001 | 46,493 | 8,807 | 7,117 |
| Mai | 39041 | 0,001 | 56,961 | 10,158 | 8,571 |
| Jun | 35081 | 0,001 | 49,600 | 10,483 | 9,925 |
| Jul | 33676 | 0,000 | 66,734 | 9,630 | 8,406 |
| Ago | 37655 | 0,001 | 58,562 | 11,123 | 9,691 |
| Set | 37448 | 0,000 | 62,232 | 12,282 | 10,456 |
| Out | 39558 | 0,000 | 59,710 | 10,197 | 8,905 |
| Nov | 37283 | 0,001 | 52,000 | 10,154 | 8,329 |
| Dez | 36260 | 0,000 | 47,151 | 9,634 | 7,801 |

Apesar dos erros máximos apresentarem valores elevados, através da análise da Tabela 4.7 com a distribuição de freqüência dos erros, pode-se concluir que a rede apresentou um bom desempenho uma vez que um número reduzido de simulações apresentou valores com erros elevados.

A Tabela 4.8 assim como o Gráfico 4.13 mostram de forma mais clara o desempenho da representação neural. Neles podemos observar que, no pior caso, 52% das estimativas efetuadas pela RNA apresentaram erro abaixo de 10%.

Há que se considerar ainda que as RNA foram treinadas com dados médios mensais de Tmin e Tmax e que, neste caso, os dados utilizados para teste são diários. É de se esperar que, durante a operação da rede, se forem fornecidos com entrada dados médios de Tmin e Tmax,

o desempenho se apresente melhor, semelhante ao apresentado no teste efetuado com os dados utilizados no treinamento.

**Tabela 4.7**
**Distribuição de freqüência dos erros**

| | Janeiro | | Fevereiro | | Março | | Abril | | Maio | | Junho | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Erro** | **Freq** | **(%)** | **Freq** | **(%)** | **Freq** | **(%)** | **Freq** | **(%)** | **Freq** | **(%)** | **Freq** | **(%)** |
| 00 \|– 05 | 19774 | 53,42 | 14010 | 40,76 | 15315 | 40,48 | 11974 | 37,96 | 13528 | 34,65 | 12494 | 35,61 |
| 05 \|– 10 | 10934 | 29,54 | 9619 | 27,98 | 10144 | 26,82 | 8446 | 26,78 | 9888 | 25,33 | 9094 | 25,92 |
| 10 \|– 15 | 4082 | 11,03 | 5433 | 15,81 | 5679 | 15,01 | 5153 | 16,34 | 6298 | 16,13 | 5464 | 15,58 |
| 15 \|– 20 | 1528 | 4,13 | 2882 | 8,38 | 3350 | 8,86 | 3199 | 10,14 | 3890 | 9,96 | 3268 | 9,32 |
| 20 \|– 25 | 489 | 1,32 | 1387 | 4,04 | 1872 | 4,95 | 1713 | 5,43 | 2492 | 6,38 | 1640 | 4,67 |
| 25 \|– 30 | 141 | 0,38 | 661 | 1,92 | 935 | 2,47 | 742 | 2,35 | 1478 | 3,79 | 948 | 2,70 |
| 30 \|– 35 | 48 | 0,13 | 291 | 0,85 | 398 | 1,05 | 276 | 0,87 | 943 | 2,42 | 653 | 1,86 |
| 35 \|– 40 | 14 | 0,04 | 81 | 0,24 | 119 | 0,31 | 35 | 0,11 | 433 | 1,11 | 579 | 1,65 |
| 40 \|– 45 | 2 | 0,01 | 7 | 0,02 | 14 | 0,04 | 5 | 0,02 | 86 | 0,22 | 549 | 1,56 |
| 45 \|– 50 | 1 | 0,00 | 2 | 0,01 | 2 | 0,01 | 1 | 0,00 | 2 | 0,01 | 392 | 1,12 |
| 50 \|– 55 | 0 | 0,00 | 0 | 0,00 | 1 | 0,00 | 0 | 0,00 | 2 | 0,01 | 0 | 0,00 |
| 55 \|– 60 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 | 1 | 0,00 | 0 | 0,00 |
| 60 \|– 65 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 |
| 65 \|– 70 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 |
| **Total** | **37013** | **100,00** | **34373** | **100,00** | **37829** | **100,00** | **31544** | **100,00** | **39041** | **100,00** | **35081** | **100,00** |

| | Julho | | Agosto | | Setembro | | Outubro | | Novembro | | Dezembro | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Erro** | **Freq** | **(%)** | **Freq** | **(%)** | **Freq** | **(%)** | **Freq** | **(%)** | **Freq** | **(%)** | **Freq** | **(%)** |
| 00 \|– 05 | 12123 | 36,00 | 12237 | 32,5 | 10597 | 28,30 | 13448 | 34,00 | 12235 | 32,82 | 12384 | 34,15 |
| 05 \|– 10 | 9028 | 26,81 | 9286 | 24,66 | 8979 | 23,98 | 10354 | 26,17 | 9727 | 26,09 | 9715 | 26,79 |
| 10 \|– 15 | 5533 | 16,43 | 6037 | 16,03 | 6504 | 17,37 | 6796 | 17,18 | 6623 | 17,76 | 6539 | 18,03 |
| 15 \|– 20 | 3184 | 9,46 | 3709 | 9,85 | 4087 | 10,91 | 3813 | 9,64 | 3920 | 10,51 | 3607 | 9,95 |
| 20 \|– 25 | 1716 | 5,10 | 2537 | 6,74 | 2669 | 7,13 | 2291 | 5,79 | 2248 | 6,03 | 2050 | 5,65 |
| 25 \|– 30 | 992 | 2,95 | 1639 | 4,35 | 1695 | 4,53 | 1321 | 3,34 | 1308 | 3,508 | 1159 | 3,20 |
| 30 \|– 35 | 538 | 1,60 | 1030 | 2,76 | 1181 | 3,15 | 686 | 1,73 | 746 | 2,00 | 537 | 1,48 |
| 35 \|– 40 | 309 | 0,92 | 636 | 1,69 | 810 | 2,16 | 393 | 0,99 | 320 | 0,86 | 198 | 0,55 |
| 40 \|– 45 | 151 | 0,45 | 362 | 0,96 | 490 | 1,31 | 225 | 0,57 | 136 | 0,37 | 63 | 0,17 |
| 45 \|– 50 | 70 | 0,29 | 121 | 0,32 | 282 | 0,75 | 120 | 0,30 | 18 | 0,05 | 8 | 0,02 |
| 50 \|– 55 | 24 | 0,07 | 47 | 0,13 | 129 | 0,34 | 78 | 0,20 | 2 | 0,01 | 0 | 0,00 |
| 55 \|– 60 | 7 | 0,02 | 14 | 0,04 | 24 | 0,06 | 33 | 0,08 | 0 | 0,00 | 0 | 0,00 |
| 60 \|– 65 | 0 | 0,00 | 0 | 0,00 | 1 | 0,00 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 |
| 65 \|– 70 | 1 | 0,00 | 0 | 0,00 | 0 | 0,00 | 0 | 0,00 | 0,00 | 0,00 | 0 | 0,00 |
| **Total** | **33676** | **100,00** | **37655** | **100,00** | **37448** | **100,00** | **39558** | **100,00** | **37283** | **100,00** | **36260** | **100,00** |

**Tabela 4.8**
**Percentual de estimativa X erro acumulado**

| Mês | Erro Acumulado (%) | | | |
|---|---|---|---|---|
| | = 5% | = 10% | = 15% | = 20% |
| Jan | 53,42 | 82,97 | 93,99 | 98,12 |
| Fev | 40,76 | 68,74 | 84,55 | 92,93 |
| Mar | 40,48 | 67,30 | 82,31 | 91,17 |
| Abr | 37,96 | 64,73 | 81,07 | 91,21 |
| Mai | 34,65 | 59,98 | 76,11 | 86,07 |
| Jun | 35,61 | 61,54 | 77,11 | 86,43 |
| Jul | 36,00 | 62,81 | 79,24 | 93,79 |
| Ago | 32,50 | 57,16 | 73,19 | 83,04 |
| Set | 28,30 | 52,28 | 69,64 | 80,56 |
| Out | 34,00 | 60,17 | 77,35 | 86,99 |
| Nov | 32,82 | 58,91 | 76,67 | 87,18 |
| Dez | 34,15 | 60,95 | 78,98 | 88,93 |

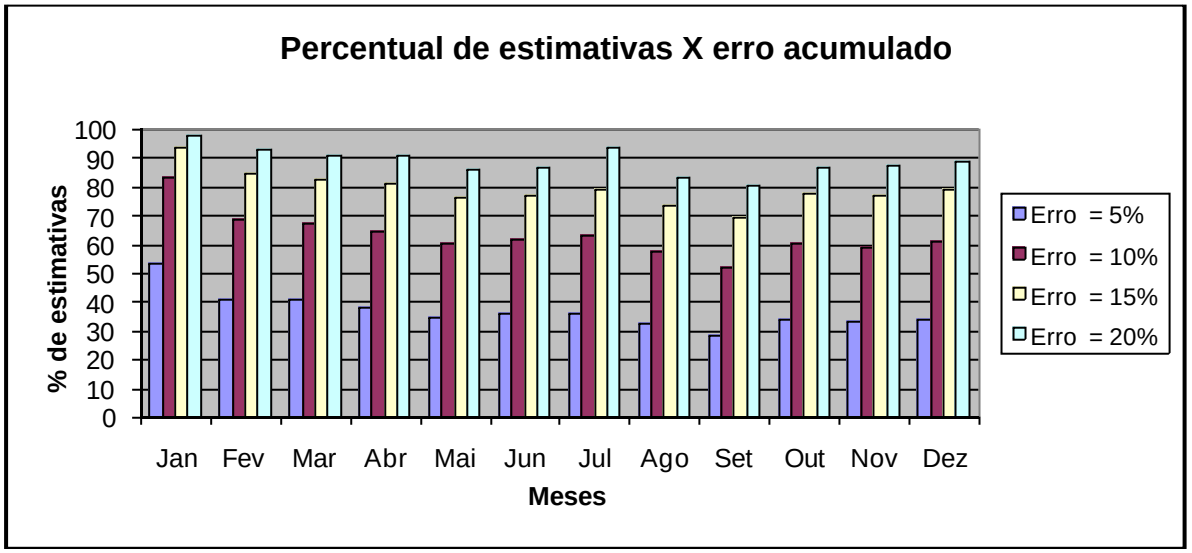


Gráfico 4.13 – Percentual de estimativa X erro acumulado

## 4.3 Conclusões

Diante da ausência de dados sobre radiação solar, vento e umidade relativa do ar, optou-se por não utilizar dados de radiação solar e vento. Diante dos possíveis impactos negativos que a não utilização de dados sobre umidade relativa do ar poderiam causar, optou-se por estimar esses dados através da utilização de redes neurais artificiais. Para essa tarefa utilizou-se um conjunto de dados históricos de 255 estações meteorológicas, observados ao

longo dos anos de 1995 a 2006. Esse conjunto de dados se mostrou suficientemente representativo.

No processo de treinamento das RNA, foram utilizadas cinco variáveis como entrada, apesar do conjunto de dados históricos conter um número maior de variáveis. Essa quantidade foi limitada pelo número de variáveis disponíveis acerca das cidades onde se necessitava estimar os dados de UR.

Após a preparação dos dados e treinamento de 12 RNA, para representar cada mês do ano, os resultados obtidos foram considerados satisfatórios.

Os testes efetuados com dados não utilizados no treinamento da rede apresentaram erros máximos elevados, porém o número de ocorrências com estes valores foi pequeno. O erro médio foi baixo, sendo no pior caso inferior a 13%.

Se, no processo de treinamento das RNA fossem utilizadas como entrada, variáveis mais intimamente relacionadas a umidade relativa, tal como precipitação atmosférica (chuva), o desempenho da rede poderia melhorar, uma vez que existe uma forte relação entre esses dois fatores.

A representação neural, tendo como parâmetros de entrada latitude, longitude, altitude, temperatura mínima e temperatura máxima, torna esse modelo útil e flexível para estimar a umidade relativa, uma vez que esses parâmetros são facilmente obtidos em qualquer estação meteorológica.

# 5. MINERAÇÃO DE DADOS

Depois de concluída a montagem do banco de dados, as próximas etapas do processo KDD, segundo Fayyad *et al*. (1996), são a mineração de dados e a interpretação/avaliação dos resultados. Neste capítulo estas etapas são descritas.

Na etapa de mineração de dados, o algoritmo *K-means* foi utilizado para a formação dos grupos.

Ao final deste capítulo é apresentada a análise dos resultados obtidos, que foi feita de acordo com um critério proposto neste trabalho e também de acordo com critérios definidos pela equipe de melhoramento genético da Embrapa Milho e Sorgo. A análise sob os diferentes critérios apresentou divergências quanto à coerência dos agrupamentos formados.

## 5.1 Aplicação da técnica de clusterização

O algoritmo *K-means* foi escolhido para ser utilizado neste trabalho por ser amplamente relatado na literatura como o método mais utilizado em análises de agrupamentos. A distância euclidiana foi utilizada como medida de similaridade.

Foi utilizada a implementação de *K-means* disponível no software Matlab, versão 6.5. O script abaixo foi utilizado para a formação dos grupos:

```
1  load bdpivotado_kdd.prn
2  [cidx, ctrs] = kmeans(bdpivotado_kdd, Num_Clusters, 'dist','sqEuclidean', 'rep',5,
   'disp','final');
```

A primeira linha carrega o banco de dados em uma matriz de 35 linhas e 224 colunas, a segunda linha executa o agrupamento. O parâmetro ***Num_Clusters*** define a quantidade de agrupamentos que se deseja formar. O parâmetro ***'dist','sqEuclidean',*** define que a distância euclidiana será utilizada como medida de similaridade. O parâmetro ***'rep',5*** define que o processo de agrupamento será repetido cinco vezes,com um conjunto inicial de centróides

diferente para cada uma delas . O parâmetro ***'disp','final'*** define que os resultados sejam exibidos somente ao final do processo.

### 5.1.1 Definição da quantidade de agrupamentos *(k)*

Conforme já dito, o *K-means* é um método de agrupamento não hierárquico ou de particionamento. Nesses métodos, o número de agrupamentos deve ser definido a priori.

O objetivo deste trabalho foi reduzir a quantidade de experimentos, no entanto, a quantidade a reduzir não havia sido definida previamente. Sendo assim, optou-se por executar diversas vezes o algoritmo de agrupamento, variando o percentual de redução de experimentos de 5% a 30%.

Dos 35 experimentos iniciais, aplicando-se uma redução de 5%, agruparam-se duas cidades, conseqüentemente a quantidade de agrupamentos formados foi de 33. A Tabela 5.1 mostra a quantidade de agrupamentos formados para cada um dos percentuais de redução utilizados.

**Tabela 5.1**
**Quantidade agrupamentos formados**

| % de redução | Número de experimentos reduzidos | Total de agrupamentos formados *(k)* |
|---|---|---|
| 5% | 2 | 33 |
| 10% | 3 | 32 |
| 15% | 5 | 30 |
| 20% | 7 | 28 |
| 25% | 9 | 26 |
| 30% | 10 | 25 |

Desta forma, o parâmetro ***Num_Clusters*** foi alterado com o valor da coluna “Total de agrupamentos formados” da Tabela 5.1, para cada vez que o algoritmo foi executado.

### 5.1.2 Visualização dos agrupamentos formados

As Tabelas 5.2 a 5.7 mostram, para cada percentual de redução, as cidades que foram agrupadas. Os agrupamentos não listados nas tabelas são formados por um único experimento.

**Tabela 5.2**
**Cidades agrupadas com redução de 5% do número de experimentos**

| % redução | Número de agrupamentos formados | Experimentos Agrupados | UF |
|---|---|---|---|
| 5% | 33 | Baixo Grande Ribeiro<br>São Raimundo das Mangabeiras | PI<br>MA |
| | | Dourados (safrinha)<br>Maracaju (safrinha) | MS<br>MS |

**Tabela 5.3**
**Cidades agrupadas com redução de 10% do número de experimentos**

| % redução | Número de agrupamentos formados | Experimentos Agrupados | UF |
|---|---|---|---|
| 10% | 32 | Baixo Grande Ribeiro<br>São Raimundo das Mangabeiras | PI<br>MA |
| | | Goiânia (safrinha)<br>Rio Verde (safrinha) | GO<br>GO |
| | | Patos de Minas<br>Uberlândia | MG<br>MG |

**Tabela 5.4**
**Cidades agrupadas com redução de 15% do número de experimentos**

| % redução | Número de agrupamentos formados | Experimentos Agrupados | UF |
|---|---|---|---|
| 15% | 30 | Dourados (safrinha)<br>Maracaju (safrinha) | MS<br>MS |
| | | Baixo Grande Ribeiro<br>São Raimundo das Mangabeiras | PI<br>MA |
| | | Brasilândia do Sul<br>Maracaju | PR<br>MS |
| | | Patos de Minas<br>Uberlândia | MG<br>MG |
| | | Goiânia (safrinha)<br>Rio Verde (safrinha) | GO<br>GO |

**Tabela 5.5**
**Cidades agrupadas com redução de 20% do número de experimentos**

| % redução | Número de agrupamentos formados | Experimentos Agrupados | UF |
|---|---|---|---|
| 20% | 28 | Baixo Grande Ribeiro | PI |
| | | Palotina | PR |
| | | São Raimundo das Mangabeiras | MA |
| | | Goiânia (safrinha) | GO |
| | | Rio Verde (safrinha) | GO |
| | | Patos de Minas | MG |
| | | Uberlândia | MG |
| | | Dourados (safrinha) | MS |
| | | Maracaju (safrinha) | MS |
| | | Campo Mourão (safrinha) | PR |
| | | Ponta Porã (safrinha) | MS |
| | | Birigui | SP |
| | | Piracicaba | SP |

**Tabela 5.6**
**Cidades agrupadas com redução de 25% do número de experimentos**

| % redução | Número de agrupamentos formados | Experimentos Agrupados | UF |
|---|---|---|---|
| 25% | 26 | Palmeiras de Goiás | GO |
| | | Rio Verde | GO |
| | | Baixo Grande Ribeiro | PI |
| | | São Raimundo das Mangabeiras | MA |
| | | Goiânia (safrinha) | GO |
| | | Rio Verde (safrinha) | GO |
| | | Birigui | SP |
| | | Piracicaba | SP |
| | | Campo Mourão (safrinha) | PR |
| | | Ponta Porã (safrinha) | MS |
| | | Campo Mourão | PR |
| | | Londrina | PR |
| | | Dourados (safrinha) | MS |
| | | Maracaju (safrinha) | MS |
| | | Brasilândia do Sul | PR |
| | | Maracaju | MS |
| | | Patos de Minas | MG |
| | | Uberlândia | MG |

**Tabela 5.7**
**Cidades agrupadas com redução de 30% do número de experimentos**

| % redução | Número de agrupamentos formados | Experimentos Agrupados | UF |
|---|---|---|---|
| 30% | 25 | Dourados (safrinha) | MS |
| | | Maracaju (safrinha) | MS |
| | | Palotina (safrinha) | PR |
| | | Birigui | SP |
| | | Piracicaba | SP |
| | | Brasilândia do Sul | PR |
| | | Maracaju | MS |
| | | Palmeiras de Goiás | GO |
| | | Rio Verde | GO |
| | | Campo Mourão | PR |
| | | Londrina | PR |
| | | Campo Mourão (safrinha) | PR |
| | | Ponta Porã (safrinha) | MS |
| | | Patos de Minas | MG |
| | | Uberlândia | MG |
| | | Baixo Grande Ribeiro | PI |
| | | São Raimundo das Mangabeiras | MA |
| | | Goiânia Safrinha | GO |
| | | Rio Verde Safrinha | GO |

### 5.1.3 Critérios para validação dos agrupamentos obtidos

Para verificar a coerência dos agrupamentos formados, foram utilizados dois critérios, sendo o primeiro deles proposto neste trabalho, e o segundo definido pela equipe de melhoramento genético da Embrapa Milho e Sorgo.

#### 5.1.3.1 Primeiro critério – validação com base na produção total do experimento

Para definição deste critério, partiu-se da seguinte hipótese:

**H1)** Se duas cidades apresentam características semelhantes de clima e solo, um experimento plantado nestas cidades, com as mesmas cultivares e sendo conduzido dentro dos mesmos critérios, deve apresentar uma produção total semelhante.

Neste critério foi proposto que, para ser considerado um agrupamento válido, os experimentos agrupados não poderiam apresentar uma variação da produção total superior a 10% ou seja, a diferença entre o experimento com menor produção e o de maior produção não poderia ser superior a 10%. A produção total foi obtida somando-se o valor da variável peso de espiga. Se duas cidades são agrupadas em um mesmo cluster e a diferença de produção entre elas é superior a 10%, este agrupamento foi considerado inadequado, portanto a eliminação de qualquer uma das cidades deste agrupamento não é recomendada. O percentual de variação foi calculado da seguinte forma:

$$Variação = (1 - \frac{Menor\ Peso\ Espiga\ \ do\ agrupamento}{Maior\ Peso\ Espiga\ \ do\ agrupamento}) \times 100$$

Nos agrupamentos com mais de dois experimentos, caso algum deles apresente diferença de produção superior a 10%, ele será excluído do agrupamento, sem, no entanto, excluir o cluster totalmente, pois os experimentos que apresentarem diferença inferior a 10% permanecerão.

As Tabelas 5.8 a 5.13 mostram a análise dos agrupamentos de acordo com este critério.

**Tabela 5.8**
**Análise de agrupamentos formados com base na produção total (Redução 5%)**

| Agrupamentos Formados | UF | Produção Total (Kg) | Variação em relação a menor e maior produção (%) | Agrupamento Válido |
|---|---|---|---|---|
| Baixo Grande Ribeiro | PI | 311008 | 0,16 | Sim |
| São Raimundo das Mangabeiras | MA | 311508 | | |
| | | | | |
| Dourados (safrinha) | MS | 293401 | 1,61 | Sim |
| Maracaju (safrinha) | MS | 288682 | | |

**Tabela 5.9**
**Análise de agrupamentos formados com base na produção total (Redução 10%)**

| Agrupamentos Formados | UF | Produção Total (Kg) | Variação em relação a menor e maior produção (%) | Agrupamento Válido |
|---|---|---|---|---|
| Baixo Grande Ribeiro | PI | 311008 | 0,16 | Sim |
| São Raimundo das Mangabeiras | MA | 311508 | | |
| | | | | |
| Goiânia (safrinha) | GO | 202753 | 22,50 | Não |
| Rio Verde (safrinha) | GO | 157133 | | |
| | | | | |
| Patos de Minas | MG | 413829 | 6,98 | Sim |
| Uberlândia | MG | 384924 | | |

**Tabela 5.10**
**Análise de agrupamentos formados com base na produção total (Redução 15%)**

| Agrupamentos Formados | UF | Produção Total (Kg) | Variação em relação a menor e maior produção (%) | Agrupamento Válido |
|---|---|---|---|---|
| Dourados (safrinha) | MS | 293401 | 1,61 | Sim |
| Maracaju (safrinha) | MS | 288682 | | |
| | | | | |
| Baixo Grande Ribeiro | PI | 311008 | 0,16 | Sim |
| São Raimundo das Mangabeiras | MA | 311508 | | |
| | | | | |
| Brasilândia do Sul | PR | 255653 | 5,37 | Sim |
| Maracaju | MS | 241921 | | |
| | | | | |
| Patos de Minas | MG | 413829 | 6,98 | Sim |
| Uberlândia | MG | 384924 | | |
| | | | | |
| Goiânia (safrinha) | GO | 202753 | 22,50 | Não |
| Rio Verde (safrinha) | GO | 157133 | | |

**Tabela 5.11**
**Análise de agrupamentos formados com base na produção total (Redução 20%)**

| Agrupamentos Formados | UF | Produção Total (Kg) | Variação em relação a menor e maior produção (%) | Agrupamento Válido |
|---|---|---|---|---|
| Baixo Grande Ribeiro | PI | 311008 | 6,71 | Sim |
| Palotina | PR | 333366 | | |
| São Raimundo das Mangabeiras | MA | 311508 | | |
| | | | | |
| Goiânia (safrinha) | GO | 202753 | 22,50 | Não |
| Rio Verde (safrinha) | GO | 157133 | | |
| | | | | |
| Patos de Minas | MG | 413829 | 6,98 | Sim |
| Uberlândia | MG | 384924 | | |
| | | | | |
| Dourados (safrinha) | MS | 293401 | 1,61 | Sim |
| Maracaju (safrinha) | MS | 288682 | | |
| | | | | |
| Campo Mourão (safrinha) | PR | 103288 | 27,53 | Não |
| Ponta Porã (safrinha) | MS | 142518 | | |
| | | | | |
| Birigui | SP | 455555 | 5,44 | Sim |
| Piracicaba | SP | 481783 | | |

**Tabela 5.12**
**Análise de agrupamentos formados com base na produção total (Redução 25%)**

| Agrupamentos Formados | UF | Produção Total (Kg) | Variação em relação a menor e maior produção (%) | Agrupamento Válido |
|---|---|---|---|---|
| Palmeiras de Goiás | GO | 292031 | 3,50 | Sim |
| Rio Verde | GO | 302631 | | |
| | | | | |
| Baixo Grande Ribeiro | PI | 311008 | 0,16 | Sim |
| São Raimundo das Mangabeiras | MA | 311508 | | |
| | | | | |
| Goiânia (safrinha) | GO | 202753 | 22,50 | Não |
| Rio Verde (safrinha) | GO | 157133 | | |
| | | | | |
| Birigui | SP | 455555 | 5,44 | Sim |
| Piracicaba | SP | 481783 | | |
| | | | | |
| Campo Mourão (safrinha) | PR | 103288 | 27,53 | Não |
| Ponta Porã (safrinha) | MS | 142518 | | |
| | | | | |
| Campo Mourão | PR | 499939 | 13,29 | Não |
| Londrina | PR | 433480 | | |
| | | | | |
| Dourados (safrinha) | MS | 293401 | 1,61 | Sim |
| Maracaju (safrinha) | MS | 288682 | | |
| | | | | |
| Brasilândia do Sul | PR | 255653 | 5,37 | Sim |
| Maracaju | MS | 241921 | | |
| | | | | |
| Patos de Minas | MG | 413829 | 6,98 | Sim |
| Uberlândia | MG | 384924 | | |

**Tabela 5.13**
**Análise de agrupamentos formados com base na produção total (Redução 30%)**

| Agrupamentos Formados | UF | Produção Total (Kg) | Variação em relação a menor e maior produção (%) | Agrupamento Válido |
|---|---|---|---|---|
| Dourados (safrinha) | MS | 293401 | 26,64 | Sim |
| Maracaju (safrinha) | MS | 288682 | 1,61 | |
| Palotina (safrinha) | PR | 215244 | 25,43 | Não (*) |
| | | | | |
| Birigui | SP | 455555 | 5,44 | Sim |
| Piracicaba | SP | 481783 | | |
| | | | | |
| Brasilândia do Sul | PR | 255653 | 5,37 | Sim |
| Maracaju | MS | 241921 | | |
| | | | | |
| Palmeiras de Goiás | GO | 292031 | 3,50 | Sim |
| Rio Verde | GO | 302631 | | |
| | | | | |
| Campo Mourão | PR | 499939 | 13,29 | Não |
| Londrina | PR | 433480 | | |
| | | | | |
| Campo Mourão (safrinha) | PR | 103288 | 27,53 | Não |
| Ponta Porã (safrinha) | MS | 142518 | | |
| | | | | |
| Patos de Minas | MG | 413829 | 6,98 | Sim |
| Uberlândia | MG | 384924 | | |
| | | | | |
| Baixo Grande Ribeiro | PI | 311008 | 0,16 | Sim |
| São Raimundo das Mangabeiras | MA | 311508 | | |
| | | | | |
| Goiânia (safrinha) | GO | 202753 | 22,50 | Não |
| Rio Verde (safrinha) | GO | 157133 | | |

(*) O agrupamento válido é formado por Dourados (safrinha) e Maracaju (safrinha). Os agrupamentos Dourados Safrina / Palotina (safrinha) e Maracaju (safrinha) e Palotina (safrinha) apresentaram variação superior a 10%

### 5.1.3.2 Segundo critério – validação com base no *ranking* de cultivares

Para a utilização deste critério, primeiro foi necessário construir o *ranking* de produção das cultivares. Esse *ranking* é formado pela cultivar e a posição de classificação da mesma, em relação a produção (peso de espiga) e é construído da seguinte forma:

1. Classificar, em ordem decrescente, a variável peso de espiga de cada experimento. Tem-se assim na primeira linha a cultivar mais produtiva e na última, a menos produtiva.
2. Montar uma tabela com a posição relativa de cada cultivar, em todos os experimentos.

Conforme já dito, na safra 2003/2004, foram instalados 35 experimentos com 32 cultivares, portanto o *ranking* ficaria conforme exemplo abaixo:

**Tabela 5.14**
**Modelo do *ranking* de produção para safra 2003/2004**

| Cultivar | Experimento 1 | Experimento 2 | … | Experimento 35 |
|---|---|---|---|---|
| 1 | posição da cultivar 1 no experimento 1 | posição da cultivar 1 no experimento 2 | | posição da cultivar 1 no experimento 35 |
| 2 | posição da cultivar 2 no experimento 1 | posição da cultivar 2 no experimento 2 | | posição da cultivar 2 no experimento 35 |
| … | … | … | … | … |
| 32 | posição da cultivar 32 no experimento 1 | posição da cultivar 32 no experimento 2 | | posição da cultivar 32 no experimento 35 |

O critério definido pela equipe de melhoramento genético da Embrapa Milho e Sorgo parte da seguinte hipótese:

**H2)** Se dois ou mais experimentos estão agrupadas e este agrupamento é coerente então a correlação entre os *rankings* destes experimentos deve ser alta.

Com base neste critério, para se verificar a coerência de um agrupamento é necessário calcular a correlação entre o *ranking* dos experimentos que compõe o agrupamento. Caso apresente uma forte correlação, esse agrupamento é considerado válido. Neste trabalho foi considerado como forte correlação valores acima de 0,7.

Considerando que o *ranking* é formado por dados não paramétricos (ordinais), é recomendada a utilização do coeficiente de correlação de Spearman (ρ), calculado da seguinte forma:

$$r = 1 - \frac{6 \times \sum_{i=1}^{n} (x - y)^2}{n^3 - n}$$ sendo:

$r$ = coeficiente de correlação de Spearman

$n$ = número de observações,

$x$,$y$ = valores observados.

O Apêndice F mostra o *ranking* com a posição de cada cultivar em todos os experimentos e o Apêndice G mostra a correlação do *ranking* de cada experimento em relação aos demais.

De acordo com este critério, nenhum dos agrupamentos formados foi considerado válido, uma vez que, em nenhum caso, a correlação entre os *rankings* foi superior a 0,7.

Com o objetivo de auxiliar a análise dos resultados através deste critério, optou-se por criar um *ranking* das correlações obtidas entre experimentos, similar ao *ranking* com posição das cultivares. Nesse *ranking* têm-se na primeira linha os experimentos com maior correlação e na última linha, o de menor correlação. Com base neste *ranking*, é possível se ter uma idéia de quão correlacionados são dois experimentos em relação aos demais. A Tabela 5.15 mostra os 20 primeiros colocados neste *ranking* e o Apêndice H são mostrados os 216 melhores. As tabelas 5.16 a 5.21 mostram os agrupamentos formados, as respectivas correlações e a posição das correlações neste *ranking*.

**Tabela 5.15**
**Correlação de Spearman – 20 maiores correlações**

| Posição | Experimento | Correlação Spearman | Posição | Experimento | Correlação Spearman |
|---|---|---|---|---|---|
| 1 | Morrinhos / Uberlândia | 0,7914 | 11 | Morrinhos / Planaltina | 0,6485 |
| 2 | Goiânia / Uberlândia | 0,7665 | 12 | Piracicaba / Uberlândia | 0,6455 |
| 3 | Montividiu (safrinha) / Uberlândia | 0,7566 | 13 | Goiânia / Piracicaba | 0,6441 |
| 4 | Morrinhos / Rio Verde (safrinha) | 0,7309 | 14 | Assis / Ponta Porã (safrinha) | 0,6419 |
| 5 | Montividiu (safrinha) / Morrinhos | 0,7265 | 15 | Campo Mourão (safrinha) / Montividiu (safrinha) | 0,6404 |
| 6 | Rio Verde (safrinha) / Uberlândia | 0,7199 | 16 | Porangatu / Rio Verde (safrinha) | 0,6393 |
| 7 | Baixo Grande Ribeiro / Porangatu | 0,7016 | 17 | Goiânia / Patos de Minas | 0,6375 |
| 8 | Goiânia / Planaltina | 0,6782 | 18 | Goiânia / Morrinhos | 0,6265 |
| 9 | Goiânia / Rio Verde (safrinha) | 0,6646 | 19 | Cristalina / Planaltina | 0,6261 |
| 10 | Piracicaba / Rio Verde (safrinha) | 0,6551 | 20 | Patos de Minas / Planaltina | 0,6250 |

**Tabela 5.16**
**Análise de agrupamentos formados com base na correlação de Spearman (Redução 5%)**

| Agrupamentos Formados | UF | Correlação | Posição no *Ranking* da Correlação | Agrupamento Válido |
|---|---|---|---|---|
| Baixo Grande Ribeiro | PI | -0,0194 | 529 | Não |
| São Raimundo das Mangabeiras | MA | | | |
| | | | | |
| Dourados (safrinha) | MS | 0,3823 | 163 | Não |
| Maracaju (safrinha) | MS | | | |

**Tabela 5.17**
**Análise de agrupamentos formados com base na correlação de Spearman (Redução 10%)**

| Agrupamentos Formados | UF | Correlação | Posição no *Ranking* da Correlação | Agrupamento Válido |
|---|---|---|---|---|
| Baixo Grande Ribeiro | PI | -0,0194 | 529 | Não |
| São Raimundo das Mangabeiras | MA | | | |
| | | | | |
| Goiânia (safrinha) | GO | 0,3970 | 142 | Não |
| Rio Verde (safrinha) | GO | | | |
| | | | | |
| Patos de Minas | MG | 0,5733 | 34 | Não |
| Uberlândia | MG | | | |

**Tabela 5.18**
**Análise de agrupamentos formados com base na correlação de Spearman (Redução 15%)**

| Agrupamentos Formados | UF | Correlação | Posição no *Ranking* da Correlação | Agrupamento Válido |
|---|---|---|---|---|
| Dourados (safrinha) | MS | 0,3823 | 163 | Não |
| Maracaju (safrinha) | MS | | | |
| | | | | |
| Baixo Grande Ribeiro | PI | -0,0194 | 529 | Não |
| São Raimundo das Mangabeiras | MA | | | |
| | | | | |
| Brasilândia do Sul | PR | 0,3416 | 198 | Não |
| Maracaju | MS | | | |
| | | | | |
| Patos de Minas | MG | 0,5733 | 34 | Não |
| Uberlândia | MG | | | |
| | | | | |
| Goiânia (safrinha) | GO | 0,3970 | 142 | Não |
| Rio Verde (safrinha) | GO | | | |

**Tabela 5.19**
**Análise de agrupamentos formados com base na correlação de Spearman (Redução 20%)**

| Agrupamentos Formados | UF | Correlação | Posição no *Ranking* da Correlação | Agrupamento Válido |
|---|---|---|---|---|
| Baixo Grande Ribeiro | PI | 0,4014 / -0,0194 | 137 / 529 | Não / Não |
| Palotina | PR | 0,0631 | 466 | Não |
| São Raimundo das Mangabeiras | MA | | | |
| | | | | |
| Goiânia (safrinha) | GO | 0,3970 | 142 | Não |
| Rio Verde (safrinha) | GO | | | |
| | | | | |
| Patos de Minas | MG | 0,5733 | 34 | Não |
| Uberlândia | MG | | | |
| | | | | |
| Dourados (safrinha) | MS | 0,3823 | 163 | Não |
| Maracaju (safrinha) | MS | | | |
| | | | | |
| Campo Mourão (safrinha) | PR | 0,1015 | 444 | Não |
| Ponta Porã (safrinha) | MS | | | |
| | | | | |
| Birigui | SP | 0,3163 | 223 | Não |
| Piracicaba | SP | | | |

**Tabela 5.20**
**Análise de agrupamentos formados com base na correlação de Spearman (Redução 25%)**

| Agrupamentos Formados | UF | Correlação | Posição no *Ranking* da Correlação | Agrupamento Válido |
|---|---|---|---|---|
| Palmeiras de Goiás | GO | 0,5227 | 57 | Não |
| Rio Verde | GO | | | |
| | | | | |
| Baixo Grande Ribeiro | PI | -0,0194 | 529 | Não |
| São Raimundo das Mangabeiras | MA | | | |
| | | | | |
| Goiânia (safrinha) | GO | 0,3970 | 142 | Não |
| Rio Verde (safrinha) | GO | | | |
| | | | | |
| Birigui | SP | 0,3163 | 223 | Não |
| Piracicaba | SP | | | |
| | | | | |
| Campo Mourão (safrinha) | PR | 0,1015 | 444 | Não |
| Ponta Porã (safrinha) | MS | | | |
| | | | | |
| Campo Mourão | PR | 0,5755 | 32 | Não |
| Londrina | PR | | | |
| | | | | |
| Dourados (safrinha) | MS | 0,3823 | 163 | Não |
| Maracaju (safrinha) | MS | | | |
| | | | | |
| Brasilândia do Sul | PR | 0,3416 | 198 | Não |
| Maracaju | MS | | | |
| | | | | |
| Patos de Minas | MG | 0,5733 | 34 | Não |
| Uberlândia | MG | | | |

**Tabela 5.21**
**Análise de agrupamentos formados com base na correlação de Spearman (Redução 30%)**

| Agrupamentos Formados | UF | Correlação | Posição no *Ranking* da Correlação | Agrupamento Válido |
|---|---|---|---|---|
| Dourados (safrinha) | MS | 0,3823 / 0,1507 | 163 / 402 | Não / Não |
| Maracaju (safrinha) | MS | 0,5312 | 54 | Não |
| Palotina (safrinha) | PR | | | |
| | | | | |
| Birigui | SP | 0,3163 | 223 | Não |
| Piracicaba | SP | | | |
| | | | | |
| Brasilândia do Sul | PR | 0,3416 | 198 | Não |
| Maracaju | MS | | | |
| | | | | |
| Palmeiras de Goiás | GO | 0,5227 | 57 | Não |
| Rio Verde | GO | | | |
| | | | | |
| Campo Mourão | PR | 0,5755 | 32 | Não |
| Londrina | PR | | | |
| | | | | |
| Campo Mourão (safrinha) | PR | 0,1015 | 444 | Não |
| Ponta Porã (safrinha) | MS | | | |
| | | | | |
| Patos de Minas | MG | 0,5733 | 34 | Não |
| Uberlândia | MG | | | |
| | | | | |
| Baixo Grande Ribeiro | PI | -0,0194 | 529 | Não |
| São Raimundo das Mangabeiras | MA | | | |
| | | | | |
| Goiânia (safrinha) | GO | 0,3970 | 142 | Não |
| Rio Verde (safrinha) | GO | | | |

Conforme pode ser visto nas Tabelas 5.16 a 5.21, a hipótese H2 é falsa para todos os agrupamentos obtidos, portanto de acordo com este critério, nenhum dos agrupamentos formados é considerado coerente.

### 5.1.4 Análise dos resultados

Considerando as seis alternativas de redução do número de experimentos propostas neste trabalho, é possível observar que à medida que se aumenta o percentual de redução, aumenta-se também a heterogeneidade dos agrupamentos. Observa-se também que, a grande maioria dos experimentos agrupados são em cidades da mesma UF, portanto geograficamente próximas, o que pode indicar condições semelhantes de clima e solo. A seguir são analisados cada um dos diferentes agrupamentos formados. A Tabela 5.22 foi utilizada para classificar a correlação existente entre os *rankings* dos experimentos agrupados.

**Tabela 5.22**
**Classificação dos coeficientes de correlação**

| Valores dos Coeficientes Calculados | Descrição |
|---|---|
| + 1,00 | Correlação positiva perfeita |
| + 0,70 a 0,99 | Correlação positiva muito forte |
| + 0,50 a 0,69 | Correlação positiva substancial |
| + 0,30 a 0,49 | Correlação positiva moderada |
| + 0,10 a 0,29 | Correlação positiva baixa |
| + 0,01 a 0,09 | Correlação positiva ínfima |
| 0,00 | Nenhuma Correlação |
| - 0,01 a 0,09 | Correlação negativa ínfima |
| - 0,10 a 0,29 | Correlação negativa baixa |
| - 0,30 a 0,49 | Correlação negativa moderada |
| - 0,50 a 0,69 | Correlação negativa substancial |
| - 0,70 a 0,99 | Correlação negativa muito forte |
| - 1,00 | Correlação negativa perfeita |

- Baixa Grande do Ribeiro / São Raimundo das Mangabeiras

  Este agrupamento está presente em todas as alternativas de redução propostas. Apesar de estarem em diferentes UF´s, as cidades são próximas geograficamente (97 Km)[6] e em altitudes (325 m / 225 m). A produção total dos experimentos foi praticamente a mesma (variação de 0,16%). O *ranking* apresentou uma correlação negativa ínfima (-0,0194). De acordo com o primeiro critério, este agrupamento é um forte candidato a ser utilizado, porém, há que se ressaltar que os dados meteorológicos utilizados foram de estações próximas (Bom Jesus e Balsas).

- Dourados (safrinha) / Maracaju (safrinha)

  Está presente em cinco das seis propostas de redução. As cidades estão na mesma UF e são próximas geograficamente (77 km) e em altitudes (430m / 384m). Observa-se também que foram agrupados experimentos de Safrinha, portanto plantados em épocas próximas. A produção total dos experimentos apresentou pequena variação (1,61%). O *ranking* apresentou uma correlação positiva moderada (0,3823). Além de apresentar uma correlação bem superior e de atender plenamente ao primeiro critério, a utilização deste agrupamento deve ser avaliada junto aos especialistas uma vez que os dados meteorológicos utilizados para ambas as cidades foram de Ponta Porã, o que certamente influenciou na sua formação.

[6] Distância geodésica

- Goiânia (safrinha) / Rio Verde (safrinha)

  Está presente em cinco das seis propostas de redução. As cidades estão na mesma UF e em altitudes próximas (749m / 715m), porém, geograficamente estão distantes (217 km). Os experimentos agrupados foram de Safrinha. O *ranking* apresentou uma correlação positiva moderada (0,3970), porém, a produção total dos experimentos apresentou variação muito além do limite estabelecido (22,5%) o que indica que este agrupamento não deve ser utilizado.

- Patos de Minas / Uberlândia

  Está presente em cinco das seis propostas de redução. As cidades estão na mesma UF e em altitudes próximas (832m / 863m). Geograficamente distam 189 km. Os dados meteorológicos utilizados para Uberlândia são de Uberaba. A variação da produção está abaixo do limite (6,98%) e o *ranking* apresentou uma correlação positiva substancial (0,5733) sendo a segunda maior correlação dentre todos os agrupamentos, portanto a utilização deste agrupamento é fortemente recomendada.

- Brasilândia do Sul / Maracaju

  Está presente em três das seis propostas de agrupamento. São de UF´s diferentes e estão a 333 km de distância. A altitude é muito próxima (378m / 384m). A variação da produção foi abaixo do limite (5,37%) e o *ranking* apresentou uma correlação positiva moderada (0,3416). Há que se ressaltar que os dados meteorológicos utilizados são de estações próximas (Umuarama e Ponta Porã), por esse motivo, sugere-se avaliar a possibilidade de utilização deste agrupamento com especialistas.

- Baixa Grande do Ribeiro / Palotina / São Raimundo das Mangabeiras

  Todas as cidades são de UF´s diferentes mas altitudes semelhantes (325m / 333m / 225m). Baixa Grande do Ribeiro e São Raimundo são próximas geograficamente (97 km) e muito distantes de Palotina (2043 km / 2015 km). A variação da produção foi abaixo do limite (6,71%). A correlação do *ranking* entre Baixa Grande e Palotina foi positiva e moderada (0,4014). Baixa Grande e São Raimundo apresentaram correlação negativa ínfima (-0,0194) e Palotina e São Raimundo, positiva ínfima (0,0631). A utilização do agrupamento Baixa Grande / São Raimundo é altamente recomendada, conforme já dito porém a inclusão de Palotina neste agrupamento deve ser melhor avaliada junto aos especialistas.

- Campo Mourão (safrinha) / Ponta Porã (safrinha)

  Situadas em UF´s diferentes e não muito próximas geograficamente (380 km). As altitudes são próximas (585m / 665m). Apresentou a maior variação dentre todos os agrupamentos (27,53%). A correlação do *ranking* foi positiva e baixa (0,1015). A utilização deste agrupamento não é recomendada.

- Palmeiras de Goiás / Rio Verde

  Situadas na mesma UF, próximas geograficamente (153 km) e em altitudes ligeiramente diferentes (596m / 715m). Os dados meteorológicos de Palmeiras de Goiás são de Varjão. Apresentou baixa variação da produção (3,5%) e correlação do *ranking* positiva e substancial (0,5227). Este agrupamento é forte candidato a ser utilizado.

- Birigui / Piracicaba

  Apesar de situados na mesma UF, não são tão próximos geograficamente (320 km) e em altitude (406m / 596m). A variação da produção foi abaixo do limite (5,44%) e a correlação do *ranking* foi positiva e moderada (0,3163). Este agrupamento está presente em três das seis propostas de redução. Os dados meteorológicos utilizados para Birigui são de Votuporanga, por isso a utilização deste agrupamento deve ser melhor avaliada junto aos especialistas.

- Campo Mourão / Londrina

  Situados na mesma UF e altitude (585m), são próximos geograficamente (149 km). Além de apresentar variação da produção superior ao limite (13,29%), este agrupamento está presente em apenas uma das propostas de redução e é formado somente com quando o maior percentual de redução é utilizado (30%) . Há que se ressaltar, entretanto que esse agrupamento apresentou uma correlação positiva e substancial, sendo a maior dentre todos os agrupamentos formados (0,5755). De acordo com o primeiro critério, a utilização deste agrupamento não é recomendada, entretanto, pelo segundo critério, é o melhor agrupamento formado.

### 5.1.5 Considerações finais

A avaliação da validade dos agrupamentos formados, feita através de dois critérios, gerou divergências, uma vez que, de acordo com o primeiro critério, de todos os agrupamentos formados, apenas quatro foram considerados não válidos. Já para o segundo critério, definido pela equipe de melhoramento genético consultada, nenhum agrupamento foi considerado válido.

Conforme já dito, o objetivo da análise de cluster é agrupar amostras semelhantes, de forma que as amostras dentro de um grupo sejam mais similares entre si e menos similares (dissimilares) com as amostras de outro grupo. Neste trabalho, as amostras a agrupar são formadas por dados de solo, clima e de todo o desenvolvimento de plantas de milho. A análise de similaridade considera, portanto, todos esse fatores.

O primeiro critério de validação dos agrupamentos é baseado na produção total dos experimentos, o que, em última instância, pode ser considerado como a síntese de todos os elementos avaliados (solo - clima - planta), uma vez que a produção é resultado da interação de todos estes elementos.

O segundo critério de avaliação, ao considerar somente o *ranking*, não leva em conta as características do solo, clima e todas as outras características das plantas avaliadas pela rede de ensaios (altura da planta, altura da espiga, etc.). De acordo com este critério, os agrupamentos mais representativos são formados pelas cidades que apresentam a maior correlação do *ranking*, não importando se essas cidades apresentam qualquer semelhança quanto ao clima e solo. Neste critério é possível formar agrupamentos com cidades com características de solo e clima completamente diferentes. A Tabela 5.22 ilustra essa situação.

Foram selecionados os 10 agrupamentos de maior correlação do *ranking* e em seguida calculou-se a similaridade para cada um dos componentes estudados (solo, clima e planta). As similaridades dos componentes foram ordenadas, do mais similar para o menos similar. Através do campo “Posição” da Tabela 5.23 é possível comparar a similaridade em relação as demais cidades.

**Tabela 5.23**
**Dez melhores agrupamentos com base no *ranking*.**

| Agrupamento | Correlação *Ranking* (1 = melhor) | | Similaridade Total (0 = melhor) | | Similaridade Plantas (0 = melhor) | | Similaridade Solos (0 = melhor) | | Similaridade Clima (0 = melhor) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Valor | Posição | Valor | Posição | Valor | Posição | Valor | Posição | Valor | Posição |
| Morrinhos / Uberlândia | 0,7914 | 1 | 2,9383 | 341 | 2,8205 | 383 | 0,2061 | 61 | 0,7974 | 79 |
| Goiânia / Uberlândia | 0,7665 | 2 | 2,2878 | 157 | 1,9995 | 219 | 0,1289 | 20 | 1,1044 | 287 |
| Montividiu (safrinha) / Uberlândia | 0,7566 | 3 | 2,0394 | 83 | 1,9325 | 188 | 0,1930 | 56 | 0,6224 | 17 |
| Morrinhos / Rio Verde (safrinha) | 0,7309 | 4 | 3,1378 | 393 | 2,9993 | 408 | 0,0496 | 7 | 0,9204 | 166 |
| Montividiu (safrinha) / Morinhos | 0,7265 | 5 | 2,6678 | 281 | 2,5485 | 330 | 0,0443 | 5 | 0,7876 | 71 |
| Rio Verde (safrinha) / Uberlândia | 0,7199 | 6 | 3,4071 | 448 | 3,3219 | 466 | 0,2302 | 77 | 0,7213 | 38 |
| Baixo Grande Ribeiro / Porangatu | 0,7016 | 7 | 2,7818 | 305 | 2,1002 | 239 | 1,3777 | 632 | 1,1957 | 373 |
| Goiânia / Planaltina | 0,6782 | 8 | 1,9198 | 54 | 1,7774 | 143 | 0,2369 | 79 | 0,6858 | 29 |
| Goiânia / Rio Verde (safrinha) | 0,6646 | 9 | 3,1807 | 400 | 2,9490 | 401 | 0,1391 | 32 | 1,1836 | 362 |
| Piracicaba / Rio Verde (safrinha) | 0,6551 | 10 | 4,4509 | 569 | 4,3476 | 548 | 0,3280 | 118 | 0,8952 | 150 |

Através da Tabela 5.22 é possível observar que, de acordo com o segundo critério, o melhor agrupamento (Morrinhos / Uberlândia) é 341º colocado em similaridade total. Observa-se também que, dentre os 10 melhores agrupamentos, o melhor em termos de similaridade ocupa a 54ª posição (Goiânia / Planaltina).

Considerando a formação de agrupamentos com base na similaridade dos elementos e analisando os dados da Tabela 5.23, pode-se concluir que o segundo critério não é o mais adequado para se avaliar a qualidade dos agrupamentos formados. Por outro lado, a análise do *ranking* parece ser o instrumento adequado para avaliar a estabilidade do comportamento das diversas cultivares avaliadas, característica essa sempre perseguida por pesquisadores em melhoramento genético.

# 6. CONCLUSÕES

Nesta dissertação foi descrita a utilização do processo de descoberta de conhecimento em banco de dados (KDD) aplicado em um conjunto de dados sobre experimentos de milho. Esses experimentos, também chamados de ensaios, são realizados em diversas cidades do Brasil e são utilizados para avaliação de cultivares de milho, antes do lançamento comercial das mesmas. Durante todo o ciclo do plantio, diversas características das plantas são monitoradas e um volume considerável de dados é gerado a partir de cada ensaio. A diversidade de condições climáticas e de solo encontradas no Brasil obriga a instalação de ensaios numa grande quantidade de cidades. Aliado a isto, a necessidade de mão de obra especializada e toda a logística envolvida na implantação e condução destes experimentos, tornam sua realização bastante onerosa.

O objetivo da utilização do processo de KDD foi reduzir o número de experimentos plantados, sem no entanto, comprometer os resultados necessários para a perfeita avaliação destas cultivares.

Foi proposto um modelo da interação da planta de milho com o meio ambiente e foram definidas as principais variáveis de cada um dos agentes envolvidos (planta, solo, clima). A influência das variáveis de solo e clima nas características observadas nas plantas de milho foi mapeada.

Em face da ausência de dados definidos no modelo e considerados imprescindíveis, foi proposta uma representação computacional neural para estimar a umidade relativa do ar, a partir de informações sobre latitude, longitude, altitude, temperatura mínima e temperatura máxima. A utilização de redes neurais artificiais para estimar a umidade relativa do ar se mostrou eficiente, conforme foi demonstrado no Capítulo 4. As RNA propostas, tendo como entrada dados facilmente obtidos em qualquer estação meteorológica, torna esse modelo útil e flexível para estimar a umidade relativa, para qualquer região do Brasil.

Utilizando técnicas de *Data Mining*, mais especificamente a análise de similaridade, foram construídos agrupamentos que, de acordo com a metodologia proposta neste trabalho, são similares em termos de solo, clima e comportamento das plantas de milho. Foram propostas seis alternativas de redução do número de experimentos instalados, sendo que a menor delas reduz em 5% o número de experimentos e a maior, em 30%. Foram propostos 11 agrupamentos diferentes.

A coerência dos agrupamentos formados foi avaliada através de dois critérios, sendo o primeiro deles, proposto nesta dissertação, baseado na produção total do experimento. O segundo critério, definido por pesquisadores em melhoramento genético da Embrapa Milho Sorgo, baseia-se no *ranking* com a posição das cultivares em relação a produção . Na avaliação pelo primeiro critério, dos 11 agrupamentos formados, quatro foram considerados incoerentes. Já pelo segundo critério, nenhum agrupamento foi considerado coerente.

Aferir a qualidade dos agrupamentos através do *ranking*, desprezando, portanto as características de solo e clima, não pareceu ser o mais adequado para avaliar agrupamentos formados através da análise de similaridade. Para se chegar a esta conclusão, foi feita uma análise dos 10 melhores agrupamentos, de acordo com o critério do *ranking*. Esta análise mostrou que esses agrupamentos são pouco ou nada similares e possivelmente nunca seriam formados através de técnicas de análise de similaridade.

A ordem que os agrupamentos foram apresentados reflete o grau de similaridade entre estes experimentos, portanto aqueles que foram agrupados primeiro, ou seja, nos menores percentuais de redução (5%), são considerados os mais similares, portanto os mais indicados para a utilização. No entanto, a opção de usar os agrupamentos aqui sugeridos e quais agrupamentos utilizar, deve ser avaliada junto com especialistas uma vez que depende de outros fatores como a conveniência de se reduzir a quantidade de experimentos em determinada região.

As principais contribuições deste trabalho são resumidas abaixo:

- Definição de um modelo da interação planta de milho – clima – solo e seus principais componentes;
- Definição de um modelo computacional para estimar a umidade relativa do ar para qualquer região do Brasil;
- Definição de um conjunto de cidades que são equivalentes em termos do clima, solo e desenvolvimento da planta de milho.

A utilização do processo de KDD.em análises de dados gerados pela pesquisa agropecuária também pode ser apontada como uma contribuição deste trabalho uma vez que, conforme já dito, sua utilização é hoje incipiente nesta área.

Além da não utilização de todas as variáveis inicialmente definidas no modelo, as principais restrições aos resultados obtidos são listadas abaixo:

- Utilização de parte dos dados simulados ao invés de dados coletados;

Parte dos dados dos ensaios e de parte dos dados de umidade relativa do ar foram simulados.

- Utilização de dados de solos obtidos em mapas de levantamento de solos ao invés de dados obtidos através de análises de solo;

  Dados obtidos em análises de solo refletem a real situação do solo onde foram realizados os ensaios. Dados obtidos em mapas de solos são menos confiáveis uma vez que estes mapas são feitos em escalas menores.

- Utilização de dados meteorológicos de estações próximas ao invés de dados da própria cidade;

  A ausência de estações meteorológicas em todas as cidades onde foram realizados os ensaios obrigou a utilização de dados de estações próximas. Algumas vezes estas estações não estavam dentro da distancia recomendada (25 km).

Diante dos impactos positivos que a redução do número de ensaios pode proporcionar, sugere-se como trabalho futuro que a metodologia aqui desenvolvida seja avaliada através de sua utilização com dados de ensaios de milho de outros anos e que os resultados sejam comparados. Sugere-se ainda que, para os próximos ensaios, as restrições citadas acima sejam eliminadas para que se possa avaliar os resultados com dados mais confiáveis. A utilização de outros algoritmos de análise de similaridade e a definição de novos critérios de validação também poderiam ser investigados em trabalhos futuros.

## REFERÊNCIAS

ABDULLAH, Ahsan; BROBST, Stephen; PERVAIZ, Ijaz; **Learning Dynamics of Pesticide Abuse through Data Mining**. Dunedin, NewZealand: Conferences in Research and Practice in Information Technology, Vol. 32, 2004

BARBOSA, Anderson Henrique; FREITAS, Marcílio Sousa da Rocha; NEVES, Francisco de Assis das Neves. **Confiabilidade estrutural utilizando o método de Monte Carlo e redes neurais**. Rev. Esc. Minas vol.58 n.3: Ouro Preto, 2005

BUCENE, Luciana Corpas; RODRIGUES, Luiz Henrique Antunes; MEIRA, Carlos Alberto Alves. **Mineração de Dados Climáticos para Previsão de Geada e Deficiência Hídrica para as Culturas do Café e da Cana-de-Açúcar para o Estado de São Paulo**. Campinas: Embrapa Informática Agropecuária, 2002. 41 p.: il. – (Documentos / Embrapa Informática Agropecuária ; 20)

BERTIS, B., JOHNSTON W. L., *et al*. **Data Mining in U.S. Corn Fields**. Proceedings of the First SIAM International Conference on Data Mining, 2001

BRAGA, A. P.; LUDERMIR, T. B.; CARVALHO, A. C. P. L. F. **Redes Neurais Artificiais**: **teoria e aplicações**. Rio de Janeiro: LTC, 2000. 262 p.

BUSSAB, Wilton de Oliveira; MIAZAKI, Édina Shizue; ANDRADE, Dalton Francisco de. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990.

CUNNINGHAM, S. J., and HOLMES, G. **Developing innovative applications in agriculture using data mining**. Department of Computer Science, University of Waikato, Hamilton, New Zealand, 2001.

DIAS FILHO, Moacyr B.; NETO, Miguel Simão; Serrão, Emanuel A.S. **Cluster Analysis for assessing the agronomic adaptation of panicum maximum Jacq. accessions**. Pesq. agropec. bras., Brasília, v.29, n.10, p.1509-1516, out. 1994

EMYGDIO, Beatriz Marti, *et al*. **Diversidade genética em cultivares locais e comerciais de feijão baseada em marcadores RAPD**. Pesq. agropec. bras., Brasília, v. 38, n. 10, p. 1165-1171, out. 2003.

EVERITT, Brian. **Cluster Analysis**. 2.ed. New York: Halsted Press, 1980. 137p.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **Artificial Intelligence**, v. 17, p. 37-54, 1996.

GUIMARÃES, Alaine Margarete. **Aplicação de computação evolucionária na mineração de dados físico-químicos da água e do solo.** 2005. 134 f. Tese – Unesp, Faculdade de Ciências Agronômicas, Botucatu.

HALDIKI, Maria; BATISTAKIS, Yannis; VAZIRGIANNIS, Michalis, **On clustering validation techiniques**. Journal of Intelligent Information System, v. 17,n 2-3, p107-145, Dec. 2001

HARMS, S., *et al.* **Data Mining in a Geospatial Support System for Drought Risk Management**. Proc. First National Conference on Digital Government Research, Los Angeles, California, pp. 9-16, 2000

HAYKIN, Simon. **Redes neurais: princípios e prática**. 2. ed. Porto Alegre:Bookman, 2001. 900 p.

HECHT-NIELSEN, R. **Theory of the backpropagation neural network**. Neural Networks, 1989. IJCNN., International Joint Conference on , vol., no., pp.593-605 vol.1, 18-22 Jun 1989

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. **Data Clustering: a review.** ACM Computing Surveys, New York, v. 31, n. 3, p. 265-323, Sept., 1999.

KANTARDZIC, Mehmed. **Data Mining: Concepts, Models, Methods and Algorithms.** John Wiley- Hoboken :Wiley-Interscience, IEEE Press, 2003.

KELLER FILHO, Thadeu; ASSAD, Eduardo Delgado; LIMA, Paulo Roberto. Schubnell de Rezende. **Regiões pluviometricamente homogêneas no Brasil**. Pesq. agropec. bras., Brasília, v.40, n.4, p.311-322, abr. 2005

KOVACS, Z. L. **Redes Neurais Artificiais: Fundamentos e Aplicações**. 1. ed. São Paulo: Collegium Cognitio, v. 1500. 158 p. 1996

LEVENBERG, Kenneth. **A Method for the Solution of Certain Non-Linear Problems in Least Squares**. The Quarterly of Applied Mathematics 2: 164–168, 1944.

MARQUARDT, Donald. **An Algorithm for Least-Squares Estimation of Nonlinear Parameters**. SIAM Journal on Applied Mathematics 11: 431–441, 1963.

McCULLAGH, J.; BLUFF, K.; HENDTLASS, T. **Envolving expert neural networks for meteorological rainfall estimations**. Los Alamitos, IEEE, 1999. p. 585-590.

MELO, Ricardo Wanke de; Fontana, Denise Cybis; Berlato, Moacir Antonio. **Indicadores de produção de soja no Rio Grande do Sul comparados ao zoneamento agrícola**. Pesq. agropec. bras., Brasília, v.39, n.12, p.1167-1175, dez. 2004

PESSOA, Alex Sandro Aguiar *et al*. **Simulação climática para a américa do sul usando redes neurais: anomalias de precipitação e temperatura do ar sazonal**. 2006. Instituto Nacional de Pesquisas Espaciais: São José dos Campos. Disponível em: http://mtc-m17.sid.inpe.br/rep-/sid.inpe.br/mtc-m17@80/2006/12.22.13.49 Acesso em: 14/11/2007

PIATETSKY-SHAPIRO, Gregory., **Knowledge discovery in real databases: A report on the IJCAI-89**. Workshop. AI Magazine, Vol. 11, No. 5, Jan. 1991, Special issue, 68-70.

PYLE, Dorian. **Data Preparation for Data Mining**, San Francisco: Morgan Kaufmann Publishers,1999. 540p.

QUEIROZ, R.A.B. ; Marar, J.F. . **Algoritmo adaptativo baseado no método Levenberg-Marquardt para treinamento de Redes Neurais PPS-Wavelet com entradas multi-dimensionais: detecção de faces humanas - um estudo de caso**. In: XII Encontro de Iniciação Científica e de Pós-Graduação do Instituto Tecnológico de Aeronáutica (XII ENCITA), 2006, São José dos Campos.

RUMELHART, D. E.; McClelland, J. L. **Parallel Distributed Processing: Explorations in the Microstruture of Cognition**, vol.1, Cambridge, MA: MIT Press, 1986.

SANTOS, Juliana Araújo, *et al*. **Caracteres epidemiológicos e uso da análise de agrupamento para resistência parcial à ferrugem da soja**. Pesq. agropec. bras., Brasília, v.42, n.3, p.443-447, mar. 2007.

SCHERTE, S. L., **Data mining and its potential use in textiles: A spinning mill**. PhD dissertation (2002): North Carolina State University.

SILVA, Anderson Francisco da; Costa, Luiz Cláudio; Sediyama,Gilberto. **Previsão da evapotranspiração de referência utilizando redes neurais**. Engenharia na Agricultura, Viçosa, MG, v.14, n.2, 93-99, Abr./Jun, 2006.

SONG, Xiaomu; FAN, Guoliang; RAO, Mahesh. **Automatic CRP Mapping Using Nonparametric Machine Learning Approaches**. 888 IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 43, NO. 4, APRIL 2005.

XAVIER, Gustavo Ribeiro, *et al*. **Variabilidade genética em acessos de caupi analisada por meio de marcadores RAPD**. Pesq. agropec. bras., Brasília, v.40, n.4, p.353-359, abr. 2005.

YANG, C., PRASHER S. O., LANDRY, J. A. **Use of artificial neural networks to recognize weeds in a corn field**. Journée d'information scientifique et technique en génie agroalimentaire,Saint-Hyacinthe QC, Canada, p. 60-65, 1999.

WITTEN, I.H. e E. FRANK. **Data Mining - Practical Machine Learning Tools and Techniques with JAVA Implementations**. São Francisco: Morgan Kaufmann Publishers, 2000.

ZAIANE, Osmar R. *et al*. **On Data clustering analysis: scalability, constraints and validation**. Edmonton Alberta, University of Alberta, 2003

ZANETTI, S. S. ; SOUSA, Elias Fernandes de ; OLIVEIRA, Vicente de Paulo Santos de ; ALMEIDA, Frederico Terra de. **Estimação da evapotranspiração de referência usando redes neurais artificiais**. In: XXXIV Congresso Brasileiro de Engenharia Agrícola, 2005, Canoas, RS. XXXIV Congresso Brasileiro de Engenharia Agrícola, 2005.

**Apêndice A - Parte da Tabela "Quantidade_Dias" com a localização das estações e quantidade de registros por ano.**

| Estação | Nome | UF | LAT | LONG | ALT | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 82022 | BOA VISTA(AEROPORTO) | RR | 28300 | -607100 | 140.00 | 97 | 233 | 296 | 310 | 340 | 359 | 347 | 321 | 321 | 325 | 338 | 148 |
| 82024 | BOA VISTA | RR | 28300 | -607100 | 140.00 | 125 | 233 | 302 | 339 | 337 | 357 | 352 | 31 | 31 | 20 | 14 | 2 |
| 82026 | TIRIOS | PA | 22900 | 555900 | 325.00 | 110 | 261 | 284 | 309 | 329 | 338 | 333 | 14 | 14 | 4 | 4 | |
| 82030 | AMAPA | RR | 28300 | -607100 | 140.00 | | | | | | | 3 | | | 1 | | |
| 82042 | CARACARAI | RR | 28300 | -607100 | 140.00 | 143 | 274 | 309 | 327 | 348 | 358 | 354 | 322 | 322 | 312 | 350 | 167 |
| 82067 | IAUARETE | AM | 3700 | 691200 | 119.56 | 105 | 156 | 159 | 189 | 171 | 6 | 23 | 23 | 23 | 16 | 6 | 2 |
| 82098 | MACAPA | AP | 349 | 5150 | 39 | 76 | 146 | 211 | 307 | 335 | 361 | 347 | 306 | 306 | 266 | 362 | 168 |
| 82099 | MACAPA (AEROPORTO) | AP | 349 | 5150 | 39 | 149 | 288 | 271 | 1 | 311 | 7 | 24 | 20 | 20 | 8 | 11 | |
| 82106 | UAUPES | AM | 3700 | 691200 | 119.56 | 121 | 252 | 291 | 315 | 346 | 357 | 352 | 304 | 304 | 257 | 343 | 165 |
| 82113 | BARCELOS | AM | 3700 | 691200 | 119.56 | 143 | 288 | 237 | 299 | 89 | 361 | 365 | 323 | 323 | 270 | 348 | 167 |
| 82141 | SOURE | PA | 22900 | 555900 | 325.00 | 132 | 231 | 280 | 310 | 336 | 359 | 312 | 364 | 364 | 358 | 363 | 169 |
| 82143 | SALINOPOLIS | PA | 22900 | 555900 | 325.00 | 38 | 18 | 62 | 80 | 236 | 340 | 337 | 313 | 313 | 221 | 262 | 135 |
| 82145 | TRACATEVA | PA | 22900 | 555900 | 325.00 | 31 | 71 | 145 | 274 | 285 | 357 | 355 | 349 | 349 | 356 | 365 | 167 |
| 82178 | OBIDOS | PA | 22900 | 555900 | 325.00 | 117 | 220 | 185 | 236 | 287 | 362 | 345 | 21 | 21 | 14 | 17 | 2 |
| 82181 | MONTE ALEGRE | PA | 22900 | 555900 | 325.00 | 120 | 238 | 280 | 319 | 340 | 361 | 354 | 355 | 355 | 345 | 365 | 167 |
| 82184 | PORTO DE MOZ | PA | 22900 | 555900 | 325.00 | 101 | 230 | 266 | 270 | 286 | 360 | 335 | 341 | 341 | 329 | 323 | 167 |
| 82191 | BELEM | PA | 22900 | 555900 | 325.00 | 138 | 286 | 324 | 331 | 351 | 363 | 362 | 21 | 21 | 15 | 17 | 2 |
| 82193 | BELEM(AEROPORTO) | PA | 22900 | 555900 | 325.00 | 144 | 289 | 322 | 331 | 351 | 353 | 351 | 314 | 314 | 334 | 345 | 168 |
| 82198 | TURIACU | MA | 17100 | -454100 | 44.00 | 74 | 150 | 230 | 283 | 319 | 359 | 332 | 350 | 350 | 358 | 354 | 167 |

## Apêndice B - Procedimentos adotados na preparação dos dados do conjunto de treinamento

Os dados históricos obtidos compreendiam o período de 01/01/1995 a 18/06/2006, armazenados em formato texto. Para cada dia existia um arquivo com os dados observados de todas as estações, para a referida data. Assim o arquivo "01011995" contém os dados observados do dia 01/01/1995, de todas as estações. Considerando um arquivo por dia, no período de 01/01/1995 a 18/06/2006 têm-se 4186 dias, portanto, 4186 arquivos.

A primeira linha de cada arquivo contém um cabeçalho com a estrutura do arquivo, conforme mostrado abaixo:

**Tabela B1**
**Estrutura do arquivo de dados diários**

| Campo | Descrição |
|---|---|
| CODIGO | Código da estação meteorológica |
| LONG | Longitude da estação |
| LAT | Latitute da estação |
| ALT | Altitude da estação |
| TMIN | Temperatura mínima |
| TMAX | Temperatura máxima |
| TEMP12 | Temperatura as 12:00 horas |
| TEMP18 | Temperatura as 18:00 horas |
| TEMP00 | Temperatura as 10:00 horas |
| TDEW12 | Temperatura do bulbo seco as 12:00 horas |
| TDEW18 | Temperatura do bulbo seco as 18:00 horas |
| TDEW00 | Temperatura do bulbo seco as 00:00 horas |
| VV12 | Velocidade do vento as 12:00 horas |
| VV18 | Velocidade do vento as 18:00 horas |
| VV00 | Velocidade do vento as 00:00 horas |
| PRESS12 | Pressão atmosférica as 12:00 horas |
| PRESS18 | Pressão atmosférica as 18:00 horas |
| PRESS00 | Pressão atmosférica as 00:00 horas |
| RADSAT | Radiação solar |
| UR | Umidade relativa do ar |

O campo CODIGO identifica a estação meteorológica a qual se referem os dados. Considerando o número de estações, cada arquivo tinha no máximo 255 linhas de dados e uma de cabeçalho.

Esta forma de armazenamento apresenta algumas desvantagens, a saber:

- Não permite ter informações anuais de uma estação;

- Não permite saber a quantidade de dias observados em um ano, de uma determinada estação.

Diante disto, foi desenvolvida a aplicação abaixo para separar os dados por estação e por ano.

```
1     para data de 01/01/1995 ate 18/06/2006
2         abrir arquivo data
3         ler linha de cabeçalho
4         enquanto não for fim de arquivo
5             ler linha de dados
6             pegar código da estação
7             abrir um arquivo com nome: código da estação + ano
8             gravar linha de dados
9             fechar arquivo nome: código da estação + ano
10        fim enquanto
11    fim para
```

Após a execução desta aplicação, os dados foram separados por estação e por ano. Considerando o número de estações (NE = 255) e o número de anos de observação (NA = 12) obteve-se 255 x 12 = 3060 arquivos.

Armazenados desta forma, pôde-se avaliar a quantidade de informação que se tinha sobre cada estação em determinado ano. Para facilitar essa análise, foi desenvolvida uma aplicação para contar o número de dias em cada ano, o que resultou na Tabela "Quantidade_Dias" contendo o nome da estação, seus dados de localização (código, cidade, estado, UF, latitude, longitude, altitude) e a quantidade de dias por ano (de 1995 a 2006).

Os dados de identificação de cada estação foram obtidos de outro arquivo onde constavam os campos código da estação, local e UF. Parte da Tabela "Quantidade_Dias" encontra-se no Apêndice A.

Neste ponto, alguns arquivos foram submetidos a uma inspeção visual a fim de identificar alguma inconsistência ou anormalidade. Foi observado que, em determinadas datas, nos valores de Tmin, Tmax e umidade relativa (UR) constava o valor 9999.99. Concluiu-se assim que esse valor representa a ausência desta informação na referida data.

Foi observado também as seguintes inconsistências:

1. Tmax < Tmin
2. Tmax > 56°c
3. Tmin < –5°c

Uma nova aplicação foi desenvolvida para a remoção das linhas com inconsistências utilizando as seguintes regras:

1. Remover linhas com Tmin > Tmax
2. Remover linhas com Tmin < -5°c ou Tmin > 45°c
3. Remover linhas com Tmax < 0°c ou Tmax > 45°c

Após a execução desta aplicação, 12858 registros com inconsistências foram removidos. Foi constatado que os dados de determinadas estações foram totalmente excluídos, caracterizando que estas estações não possuíam dados confiáveis. O número de estações foi portanto, reduzido.

Mesmo com a redução do número de estações e com a exclusão de registros com erro ou suspeitos, a quantidade de registros totalizava 648104. Esse volume de dados é de tal ordem que se fez necessário a segmentação dos mesmos, pelos motivos descritos adiante.

Optou-se por treinar 12 RNA, sendo uma para cada mês do ano. Para isso foi desenvolvida uma aplicação para separar os dados históricos em meses. Essa aplicação resultou em 12 arquivos denominados “janeiro”, “fevereiro”,... “dezembro”, respectivamente, um para cada mês do ano, contento em cada arquivo, os dados relativos a cada mês, de todas as estações e de todos os anos. Assim o arquivo de janeiro continha os dados do mês 01, das 255 estações, dos 12 anos.

Uma nova aplicação foi desenvolvida para contar, a partir dos arquivos “janeiro”, “fevereiro”,... “dezembro”, o número de registros válidos (sem o valor 9999.99) para Tmin, Tmax e UR, por mês, resultando no arquivo “TotalRegValido” com a seguinte estrutura:

**Tabela B2**
**Estrutura do arquivos “TotalREgValido”**

| Campo | Descrição |
|---|---|
| CODIGO | Código da estação meteorológica |
| MÊS | Número do mês |
| QT_REG | Quantidade total de registros; |
| QT_TMIN | Quantidade de registros válidos para Tmin; |
| QT_TMAX | Quantidade de registros válidos para Tmax; |
| QT_UR | Quantidade de registros válidos para UR. |

Considerando o número máximo de registros possíveis, para cada mês, chegar-se-ia a um número ainda muito grande, conforme demonstrado abaixo:

31 dias X 12 anos X 255 estações = 94860 registros

Ainda que pouco provável que algum mês tivesse essa quantidade de registros, esse número ou mesmo a metade dele, ainda era elevado e poderia inviabilizar o treinamento da RNA pelo custo computacional requerido. Além do custo computacional, essa quantidade de registros poderia levar a uma superespecialização da RNA nos dados utilizados no treinamento, comprometendo sua capacidade de generalização ou o reconhecimento de padrões não vistos no treinamento. Esse fenômeno é conhecido como *overfitting* (Kantardzic, 2003).

A fim de reduzir a quantidade de registros, optou-se por trabalhar com a média aritmética dos dados, da seguinte forma:

Foi calculada a média para cada dia do ano, durante os 12 anos. Para esta etapa foi desenvolvida uma aplicação que a partir dos arquivos “janeiro”, “fevereiro”,... “dezembro”, calculou a média diária para Tmin, Tmax e UR da seguinte forma:

Media Tmin dia 01/01 = (Tmin 01/01/1995 + Tmin 01/01/1996 + ... Tmin 01/01/2006) / 12
Media Tmin dia 02/01 = (Tmin 02/01/1995 + Tmin 02/01/1996 + ... Tmin 02/01/2006) / 12
.
.
Media Tmin dia 31/12 = (Tmin 31/12/1995 + Tmin 31/12/1996 + ... Tmin 31/12/2006) / 12

Ao final desta aplicação obteve-se 12 arquivos com nomes “MediaDiariaJan”, “MediaDiariaFev”,...,“MediaDiariaDez” contendo a média diária de todas as estações, com a seguinte estrutura:

**Tabela B3**
**Estrutura do arquivo “MediaDiariaMês”**

| Campo | Descrição |
|---|---|
| CODIGO | Código da estação meteorológica |
| LAT | Latitude |
| LONG | Longitude |
| ALT | Altitude |
| DIA_MES | Dia e mês no formato “ddmm” |
| MED_TMIN_D | Média diária de TMin |
| MED_TMAX_D | Média diária de Tmax |
| MED_UR_D | Média diária de UR |

A seguir, utilizando os arquivos gerados no passo anterior (“MediaDiariaJan”, “MediaDiariaFev”,...,“MediaDiariaDez”), calculou-se a média mensal de Tmin, Tmax, UR, para cada estação, da seguinte forma:

$$Média\_Tmin = \frac{\sum_{i=1}^{n} Tmin_i}{n}$$

$$Média\_Tmax = \frac{\sum_{i=1}^{n} Tmax_i}{n}$$

$$Média\_UR = \frac{\sum_{i=1}^{n} UR_i}{n}$$

Para tal tarefa também foi desenvolvida uma aplicação que ao final resultou no arquivo denominado “MediaMensal” com a seguinte estrutura:

**Tabela B4**
**Estrutura do arquivo “MediaMensal”**

| Campo | Descrição |
|---|---|
| CODIGO | Código da estação meteorológica |
| LAT | Latitude |
| LONG | Longitude |
| ALT | Altitude |
| MES | Mês no formato “mm” |
| MED_TMIN_M | Média mensal de TMin |
| MED_TMAX_M | Média mensal de Tmax |
| MED_UR_M | Média mensal de UR |

Optou-se por gerar um arquivo para centralizar os dados neste arquivo e assim facilitar o trabalho.

A esse arquivo foi mesclado o arquivo com a quantidade total de registros e a contagem de registros válidos por mês gerado anteriormente (arquivo “TotRegValido”), resultando em um arquivo com a seguinte estrutura:

**Tabela B5**
**Estrutura do arquivo “TotRegValidos”**

| Campo | Descrição |
|---|---|
| CODIGO | Código da estação meteorológica |
| LAT | Latitude |
| LONG | Longitude |
| ALT | Altitude |

**Tabela B5**
**Estrutura do arquivo "TotRegValidos"**

| Campo | Descrição |
|---|---|
| MÊS | Mês no formato "mm" |
| MED_TMIN_M | Média mensal de TMin |
| MED_TMAX_M | Média mensal de Tmax |
| MED_UR_M | Média mensal de UR |
| QT_REG | Quantidade total de registros |
| QT_REGV_TMIN | Quantidade de registros válidos para Tmin |
| QT_REGV_TMAX | Quantidade de registros válidos para Tmax |
| QT_REGV_UR | Quantidade de registros válidos para UR |

Com os arquivos já trabalhados e filtrados, sabia-se a quantidade de registros que se tinha sobre cada mês e quantidade de registros válidos. Neste ponto surgiu a seguinte dúvida:

- A quantidade de registros válidos é suficiente para representar cada mês?

Para responder a essa pergunta foi necessário recorrer às técnicas de estatística para determinar o "tamanho da amostra" e assim comparar com quantidade de registros válidos. Das diversas formas de se calcular o tamanho da amostra, a seguinte fórmula foi utilizada:

$$n = \frac{N \times (z \times S(x))^2}{(N-1)^2 \times e^2 + (z \times S(x))^2} \qquad \text{(equação B.1)}$$

Onde:

n = tamanho da amostra

N = tamanho da população

e = erro máximo tolerável

S(x) = desvio padrão da amostra

Foi convencionado utilizar todos os meses com 30 dias, portanto, temos:

n = quantidade de registros válidos

N = 360 (pois N = n° de dias no mês X quantidade de anos = 30 X 12 = 360)

e = 0,45 (considerando erro máximo de 10%)

z = 2,054

S(x)= era necessário calcular para Tmin, Tmax e UR para cada mês

O cálculo do desvio padrão foi feito através de uma aplicação desenvolvida para essa tarefa. Com base nos arquivos “janeiro”, “fevereiro”,..., “dezembro”, foi calculado o desvio padrão de Tmin, Tmax, UR, de cada estação, mês a mês e o resultado foi acrescentado ao arquivo “MendiaMensal”. Ao fim da execução desta aplicação, o arquivo “MediaMensal” possuía a seguinte estrutura:

**Tabela B6**
**Nova estrutura do arquivo “MediaMensal”**

| Campo | Descrição |
|---|---|
| CODIGO | Código da estação meteorológica |
| LAT | Latitude |
| LONG | Longitude |
| ALT | Altitude |
| MÊS | Mês no formato “mm” |
| MED_TMIN_M | Média mensal de TMin |
| MED_TMAX_M | Média mensal de Tmax |
| MED_UR_M | Média mensal de UR |
| QT_REG | Quantidade total de registros |
| QT_REGV_TMIN | Quantidade de registros válidos para Tmin |
| QT_REGV_TMAX | Quantidade de registros válidos para Tmax |
| QT_REGV_UR | Quantidade de registros válidos para UR |
| DESVIO_PAD_TMIN | Desvio padrão Tmin |
| DESVIO_PAD_TMAX | Desvio padrão Tmax |
| DESVIO_PAD_UR | Desvio padrão UR |

A seguir esse arquivo foi importado para uma planilha Excel e três novas colunas foram calculadas, com base na equação B.1, citada acima:

- Tamanho da amostra Tmin
- Tamanho da amostra Tmax
- Tamanho da amostra UR

De posse das informações sobre o tamanho da amostra e da quantidade de registros válidos pôde-se verificar se as amostras disponíveis eram suficientes para representar cada mês. Assim mais três colunas foram acrescentadas na planilha:

- Amostra Tmin Ok
- Amostra Tmax Ok
- Amostra UR Ok

Com o seguinte conteúdo:

- *Se quantidade de registros válidos Tmin ≥ Tamanho amostra Tmim verdadeiro senão falso*
- *Se quantidade de registros válidos Tmax ≥ Tamanho amostra Tmax verdadeiro senão falso*
- *Se quantidade de registros válidos UR ≥ UR Tamanho amostra verdadeiro senão falso*

Para ser utilizada, uma amostra deveria ser "verdadeira" para Tmin, Tmax e UR. Se o valor de algum dos três fosse falso, essa amostra não seria utilizada. Para facilitar a busca por amostras válidas a coluna "Usar Mês" foi acrescentada com o seguinte conteúdo:

- *Se (amostraTminOk = "verdadeiro" e amostraTmaxOk = "verdadeiro" e amostraUR = "verdadeiro" então "verdadeiro" senão "falso")*

Através desta coluna pôde-se separar somente as estações meteorológicas consideradas representativas. Somente estas estações foram utilizadas no treinamento das RNA. As estações consideradas não representativas foram excluídas desta planilha.
Parte desta tabela que foi denominada "ResumoGeral" se encontra no Apêndice C

**Resumo das etapas de preparação dos dados do conjunto de treinamento**

A tabela a seguir mostra um resumo de todas as etapas da preparação do conjunto de dados utilizado no treinamento das RNA.

**Tabela B7**
**Resumo das etapas de preparação dos dados do conjunto de treinamento**

| Etapa | Tarefa | Entrada | Saída |
|---|---|---|---|
| 1 | Separar arquivos por estação e ano | Arquivos originais armazenados por data. Cada dia um arquivo total 4188 arquivos | Arquivos separados por estação e por ano sendo 255 estações e 12 anos totalizando 3060 arquivos **(Saída 1)** Nome dos arquivos: CodEstação + Ano |
| 2 | Contar quantidade de dias por estação e por ano que têm dados | Arquivos de saída da etapa 1 **(Saída 1)** | Tabela com quantidade de dias por estação e por ano **(Tabela "Quantidade_Dias" - Apêndice A)** |
| 3 | Remoção de inconsistencias (Tmin >Tmax, Tmin < -5°c ou Tmin > 45°c, Tmax < 0°c ou Tmax > 45°c), | Arquivos de saída da etapa 1 **(Saída 1)** | Arquivos da Saída 1 sem registros inconsistentes (12858 registros a menos) Nome dos arquivos: E_+CodEstação+Ano **(Saída 2)** |

**Tabela B7**
**Resumo das etapas de preparação dos dados do conjunto de treinamento**

| Etapa | Tarefa | Entrada | Saída |
|---|---|---|---|
| 4 | Separação de arquivos em meses | Arquivos de saída da etapa 3 **(Saída 2)** | 12 arquivos denominados ("janeiro", "fevereiro"....) contendo dados do respectivo mês **(Saída 3)** |
| 5 | Contar número de registros válidos por mês de Tmin, Tmax e UR | Arquivos de saída da etapa 4 ("janeiro", Fevereiro" ...."dezembro") **(Saída 3)** | Arquivo "TotRegValido" contendo quantidade de registro válidos por estação e por mês de Tmin, Tmax, UR **(Saída 4)** |
| 6 | Calcular Média diária de Tmin, Tmax, UR | Arquivos de saída da etapa 4 ("janeiro", Fevereiro" ...."dezembro") **(Saída 3)** | 12 arquivos denominados "MediaDiariaJan", "MediaDiariaFev", ..., "MediaDiariaDez" **(Saída 5)** |
| 7 | Calcular Média Mensal de Tmin, Tmax, UR | Arquivos de saída da etapa 6 ("MediaDiariaJan" ,..., "MediaDiariaDez") **(Saída 5)** | Arquivo "MediaMensal" **(Saída 6)** |
| 8 | Mesclar arquivo com quantidade de registros válidos por mês | Arquivo de saída da etapa 5 ("TotRegValido") **(Saída 4)** e arquivo "MediaMensal" **(Saída 6)** | Arquivo "MediaMensal" acrescido de colunas de contagem de registros válidos de Tmin, Tmax, UR **(Saída 7)** |
| 9 | Calcular desvio padrão de Tmin, Tmax, UR | Arquivos "janeiro", "fevereiro"... **(Saída 3)** e arquivo "MediaMensal" **(Saída 7)** | Arquivo "MediaMensal" acrescido de colunas Desvio Tmin, Desvio Tmax, Desvio UR **(Saída 8)** |
| 10 | Importar para o Excel o arquivo "MediaMensal" e calcular "Tamanho Amostra Tmin", "Tamanho Amostra Tmax", "Tamanho Amostra UR" | Arquivo "MediaMensal" **(Saída 8)** | Planilha Excel "ResumoGeral" **(Saída 9)** |
| 11 | Inserir na tabela "ResumoGeral" colunas "Amostra Tmin Ok", "Amostra Tmax Ok", "Amostra UR Ok" | Planilha Excel "ResumoGeral" **(Saída 9)** | Planilha Excel "ResumoGeral" acrescida das colunas "Amostra Tmin Ok", "Amostra Tmax Ok", "Amostra UR Ok" **(Saída 10)** |
| 12 | Inserir na planilha ResumoGeral coluna "Usar Mês" | PlanilhaExcel "Resumo Geral" **(Saída 10)** | Planilha Excel "ResumoGeral" acrescida da coluna "Usar Mês" **(Saída 11)** |

**Apêndice C - Parte da tabela "ResumoGeral" contendo a quantidade amostras necessárias para representar cada mês de uma estação.**

| | | | N= 360 | | z= 2,054 | | e= 0,45 | | | N (umi)= 360 | z (umi)= 2,054 | | e (umi)= 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Estacao** | **Mês** | **Qt_Tmin** | **Qt_Tmax** | **Qt_Umi** | **Desvio_Tmin** | **Desvio_Tmax** | **Desvio_Umi** | **Amostra Tmin** | **Amostra Tmax** | **Amostra Umidade** | **Tmin OK** | **Tmax OK** | **Umi Ok** | **Usar Mês** |
| E_82022 | 1 | 250 | 220 | 281 | 0,87 | 1,6 | 8,93 | 15 | 47 | 3 | Sim | Sim | Sim | VERDADEIRO |
| E_82022 | 2 | 246 | 216 | 275 | 0,95 | 1,61 | 8,61 | 18 | 47 | 3 | Sim | Sim | Sim | VERDADEIRO |
| E_82022 | 3 | 250 | 238 | 301 | 1,18 | 1,65 | 9,6 | 27 | 49 | 4 | Sim | Sim | Sim | VERDADEIRO |
| E_82022 | 4 | 224 | 198 | 281 | 1,16 | 2,26 | 11,9 | 26 | 82 | 6 | Sim | Sim | Sim | VERDADEIRO |
| E_82022 | 5 | 288 | 244 | 311 | 1,05 | 2,37 | 9,04 | 22 | 89 | 3 | Sim | Sim | Sim | VERDADEIRO |
| E_82022 | 6 | 230 | 216 | 268 | 0,91 | 2,06 | 7,7 | 17 | 71 | 2 | Sim | Sim | Sim | VERDADEIRO |
| E_82022 | 7 | 236 | 225 | 275 | 1,12 | 1,92 | 7,4 | 24 | 63 | 2 | Sim | Sim | Sim | VERDADEIRO |
| E_82022 | 8 | 260 | 245 | 299 | 1,15 | 2,4 | 9,05 | 26 | 90 | 3 | Sim | Sim | Sim | VERDADEIRO |
| E_82022 | 9 | 231 | 213 | 261 | 0,98 | 1,98 | 8,65 | 19 | 67 | 3 | Sim | Sim | Sim | VERDADEIRO |
| E_82022 | 10 | 241 | 225 | 279 | 1,04 | 1,95 | 9,58 | 21 | 65 | 4 | Sim | Sim | Sim | VERDADEIRO |
| E_82022 | 11 | 204 | 184 | 251 | 0,99 | 1,94 | 9,85 | 19 | 65 | 4 | Sim | Sim | Sim | VERDADEIRO |
| E_82022 | 12 | 214 | 205 | 254 | 1,23 | 1,58 | 8,79 | 29 | 46 | 3 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 1 | 125 | 98 | 152 | 1,05 | 1,54 | 16,49 | 22 | 44 | 11 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 2 | 141 | 115 | 163 | 1,38 | 2,26 | 16,56 | 36 | 82 | 11 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 3 | 153 | 106 | 177 | 1,04 | 2,18 | 17,42 | 21 | 78 | 12 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 4 | 125 | 79 | 163 | 1,15 | 2,03 | 19,22 | 26 | 69 | 15 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 5 | 144 | 121 | 163 | 1,15 | 1,87 | 14,73 | 26 | 61 | 9 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 6 | 129 | 109 | 165 | 1,12 | 1,63 | 11,39 | 24 | 48 | 5 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 7 | 141 | 120 | 165 | 1,03 | 2,23 | 12,12 | 21 | 81 | 6 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 8 | 163 | 160 | 186 | 1,13 | 1,68 | 13,74 | 25 | 51 | 8 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 9 | 172 | 147 | 188 | 1,11 | 1,68 | 13,67 | 24 | 51 | 8 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 10 | 158 | 141 | 173 | 1,08 | 1,81 | 16,09 | 23 | 58 | 11 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 11 | 138 | 132 | 173 | 1,19 | 2,47 | 15,74 | 27 | 94 | 10 | Sim | Sim | Sim | VERDADEIRO |
| E_82024 | 12 | 143 | 118 | 174 | 1,05 | 2,07 | 14,15 | 22 | 72 | 8 | Sim | Sim | Sim | VERDADEIRO |
| E_82026 | 1 | 102 | 0 | 136 | 1,1 | 0 | 19 | 24 | 0 | 15 | Sim | | Sim | FALSO |
| E_82026 | 2 | 88 | 0 | 135 | 0,87 | 0 | 19,89 | 15 | 0 | 16 | Sim | | Sim | FALSO |
| E_82026 | 3 | 134 | 0 | 176 | 2,27 | 0 | 18,98 | 83 | 0 | 15 | Sim | | Sim | FALSO |

**Apêndice D - Script Matlab “treina_rna” utilizado no treinamento das redes neurais.**

```
% ------  treina_rna -----
clear; % limpa a memoria
nentradas = 5; % numero de entradas
neucam_esc = 11; % numero de neurônios da camada escondida
neucam_sai = 1; % numero de neurônios da camada de saída
load mes11.real; % carregando os conjuntos de treinamento já normalizados
mes = mes11; % MUDAR NOME DO ARQUIVO
[Linha, Col] = size(mes);
Entrada = mes(1:Linha,1:5); % carregando a entrada
Saida = mes(1:Linha,6); % carregando a saida
[En,mine,maxe] = premnmx(Entrada');% normaliza transposto das entradas
[Sn,mins,maxs] = premnmx(Saida');% normaliza transposto da saida
Ent = En; % Entradas
Sai = Sn; % Saída
% cria a rede neural ----------------------------------------------------
net = newff(minmax(Ent),...
          [neucam_esc, neucam_sai],...
          {'tansig','tansig'},...
          'trainlm');
% ---------------------------------------------------------------------
% inicializa a rede neural ----------------------------------------------
net = init(net);
% ---------------------------------------------------------------------
net.trainParam.goal = 0.001; % erro
net.trainParam.show = 50; % tempo, em iterações, para display na tela
net.trainParam.epochs = 3000; % numero de iterações
% treina a rede neural --------------------------------------------------
[net,tr,Y,E] = train(net,Ent,Sai);
% ---------------------------------------------------------------------

Limites = minmax(Ent);
w1 = net.iw{1,1}; % pesos camada escondida
w2 = net.lw{2,1}; % pesos camada de saida
b1 = net.b{1}; % bias camada escondida
b2 = net.b{2}; % bias camada de saida

% MUDAR NOME DO ARQUIVO-------------------------------------------------
save pebili11 mins maxs Limites w1 w2 b1 b2; % salva os pesos e bias da rna treinada na
variável 'pesosbias'

SaidaRealRna = postmnmx(Y,mins,maxs);
Compara = [mes(:,6)'; SaidaRealRna; abs(mes(:,6)'-SaidaRealRna)];
MaxErro = max(abs(mes(:,6)'-SaidaRealRna));
MinErro = min(abs(mes(:,6)'-SaidaRealRna));
```

**Apêndice E - Script Matlab “opera_rna” utilizado na operação das redes neurais (após o treinamento)**.

```
% ------  opera_rna -----

clc;
clear;
nentradas = 5; % número de entradas
neucam_esc = 11; % numero de neurônios da camada escondida
neucam_sai = 1; % número de neurônios da camada de saída

load pebili11.mat;
load mes11.real;
testereal = mes11(:,1:5);

[En,mine,maxe] = premnmx(testereal');% normaliza transposto das entradas

net = newff(Limites,...
   [neucam_esc, neucam_sai],...
   {'tansig','tansig'});

net = init(net);

net.iw{1,1} = w1;
net.lw{2,1} = w2;
net.b{1} = b1;
net.b{2} = b2;

% Simula a Rede Neural ----------------------------------------------------
SaidaSim = sim(net,En);
%-------------------------------------------------------------------------

SaidaRealSim = postmnmx(SaidaSim,mins,maxs);
```

## Apêndice F – *Ranking* com a posição de Classificação das Cultivares

| Cultivar | Assis (safrinha) | Baixo Grande Ribeiro | Birigui | Brasilândia do Sul | Campo Mourão | Campo Mourão (safrinha) | Cristalina | Dourados (safrinha) | Goianésia | Goiânia | Goiânia (safrinha) | Ipameri | Itumbiara | Londrina | Londrina (safrinha) | Maracaju |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 26 | 32 | 15 | 20 | 27 | 19 | 12 | 26 | 22 | 21 | 10 | 1 | 18 | 3 | 19 | 32 |
| 2 | 9 | 18 | 19 | 27 | 4 | 5 | 7 | 2 | 3 | 7 | 22 | 7 | 13 | 6 | 30 | 15 |
| 3 | 18 | 27 | 29 | 23 | 25 | 1 | 32 | 19 | 28 | 28 | 26 | 23 | 25 | 28 | 5 | 25 |
| 4 | 28 | 3 | 24 | 19 | 14 | 24 | 29 | 20 | 5 | 1 | 18 | 18 | 7 | 11 | 20 | 14 |
| 5 | 25 | 22 | 31 | 30 | 28 | 3 | 30 | 28 | 14 | 8 | 6 | 8 | 8 | 25 | 2 | 21 |
| 6 | 20 | 21 | 12 | 15 | 22 | 6 | 21 | 4 | 7 | 20 | 16 | 5 | 4 | 27 | 6 | 18 |
| 7 | 3 | 16 | 13 | 17 | 26 | 25 | 14 | 32 | 20 | 16 | 12 | 9 | 17 | 24 | 16 | 20 |
| 8 | 22 | 9 | 5 | 14 | 19 | 14 | 11 | 23 | 18 | 26 | 11 | 4 | 29 | 31 | 13 | 16 |
| 9 | 4 | 20 | 27 | 29 | 23 | 4 | 17 | 8 | 11 | 19 | 5 | 2 | 22 | 15 | 15 | 6 |
| 10 | 14 | 24 | 16 | 10 | 13 | 16 | 2 | 12 | 15 | 15 | 2 | 10 | 2 | 29 | 10 | 22 |
| 11 | 23 | 26 | 28 | 24 | 32 | 27 | 26 | 24 | 24 | 31 | 23 | 28 | 24 | 21 | 27 | 17 |
| 12 | 2 | 28 | 8 | 13 | 1 | 13 | 6 | 14 | 8 | 12 | 25 | 24 | 15 | 5 | 23 | 8 |
| 13 | 8 | 2 | 6 | 11 | 6 | 9 | 13 | 22 | 21 | 13 | 15 | 6 | 26 | 2 | 3 | 1 |
| 14 | 21 | 12 | 26 | 5 | 16 | 32 | 18 | 11 | 27 | 29 | 27 | 16 | 21 | 19 | 7 | 27 |
| 15 | 29 | 11 | 18 | 16 | 3 | 31 | 20 | 1 | 25 | 25 | 29 | 19 | 27 | 12 | 28 | 10 |
| 16 | 16 | 17 | 2 | 12 | 29 | 20 | 24 | 29 | 23 | 24 | 19 | 26 | 28 | 26 | 1 | 19 |
| 17 | 12 | 6 | 9 | 18 | 7 | 8 | 15 | 27 | 13 | 5 | 4 | 15 | 19 | 17 | 11 | 5 |
| 18 | 13 | 1 | 3 | 25 | 12 | 2 | 3 | 15 | 2 | 11 | 7 | 13 | 1 | 13 | 32 | 13 |
| 19 | 17 | 8 | 17 | 4 | 15 | 29 | 27 | 9 | 16 | 27 | 30 | 14 | 16 | 20 | 24 | 29 |
| 20 | 19 | 4 | 10 | 28 | 5 | 10 | 1 | 18 | 29 | 3 | 1 | 31 | 3 | 16 | 26 | 30 |
| 21 | 31 | 30 | 20 | 31 | 18 | 28 | 22 | 21 | 31 | 22 | 32 | 21 | 23 | 22 | 29 | 31 |
| 22 | 6 | 19 | 4 | 7 | 11 | 26 | 19 | 7 | 30 | 23 | 31 | 17 | 9 | 23 | 9 | 28 |
| 23 | 7 | 25 | 11 | 22 | 9 | 15 | 9 | 16 | 6 | 14 | 21 | 30 | 32 | 9 | 4 | 23 |
| 24 | 15 | 10 | 30 | 3 | 8 | 18 | 31 | 13 | 4 | 9 | 17 | 22 | 30 | 4 | 18 | 12 |
| 25 | 1 | 15 | 25 | 6 | 2 | 12 | 4 | 6 | 26 | 2 | 28 | 29 | 5 | 8 | 17 | 11 |
| 26 | 27 | 29 | 32 | 9 | 24 | 23 | 5 | 17 | 19 | 4 | 9 | 27 | 6 | 18 | 14 | 2 |
| 27 | 24 | 13 | 22 | 26 | 31 | 17 | 23 | 30 | 9 | 30 | 3 | 11 | 14 | 32 | 22 | 26 |
| 28 | 32 | 23 | 7 | 8 | 20 | 22 | 8 | 10 | 17 | 6 | 24 | 3 | 12 | 1 | 8 | 7 |
| 29 | 11 | 5 | 14 | 2 | 10 | 7 | 25 | 3 | 10 | 10 | 8 | 20 | 20 | 7 | 31 | 4 |
| 30 | 5 | 14 | 21 | 21 | 17 | 11 | 16 | 5 | 12 | 17 | 20 | 25 | 11 | 10 | 25 | 9 |
| 31 | 30 | 31 | 23 | 32 | 30 | 21 | 28 | 31 | 32 | 32 | 14 | 32 | 10 | 30 | 12 | 24 |
| 32 | 10 | 7 | 1 | 1 | 21 | 30 | 10 | 25 | 1 | 18 | 13 | 12 | 31 | 14 | 21 | 3 |

## Apêndice F – *Ranking* com a posição de Classificação das Cultivares

| Cultivar | Maracaju (safrinha) | Montividiu (safrinha) | Morrinhos | Palmeiras de Goiás | Palotina | Palotina (safrinha) | Patos de Minas | Piracicaba | Planaltina | Ponta Porã | Ponta Porã (safrinha) | Porangatu | Rio Verde | Rio Verde (safrinha) | São Raimundo das Mangabeiras | Sete Lagoas | Sete Lagoas - Cerrado | Sooretama | Uberlândia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 21 | 17 | 5 | 26 | 24 | 1 | 32 | 8 | 12 | 27 | 31 | 18 | 18 | 18 | 9 | 19 | 30 | 31 | 19 |
| 2 | 1 | 3 | 3 | 7 | 28 | 7 | 4 | 7 | 6 | 9 | 7 | 4 | 5 | 19 | 12 | 24 | 3 | 6 | 4 |
| 3 | 24 | 6 | 17 | 29 | 31 | 4 | 30 | 22 | 22 | 30 | 25 | 30 | 26 | 24 | 5 | 20 | 29 | 32 | 20 |
| 4 | 9 | 2 | 1 | 4 | 23 | 6 | 1 | 16 | 7 | 19 | 24 | 3 | 27 | 7 | 10 | 8 | 15 | 3 | 1 |
| 5 | 18 | 8 | 10 | 11 | 17 | 8 | 29 | 21 | 27 | 16 | 28 | 14 | 7 | 10 | 32 | 3 | 28 | 7 | 10 |
| 6 | 4 | 7 | 11 | 15 | 15 | 5 | 14 | 20 | 8 | 21 | 13 | 24 | 2 | 12 | 20 | 5 | 23 | 13 | 13 |
| 7 | 8 | 10 | 13 | 16 | 18 | 10 | 15 | 12 | 19 | 28 | 12 | 26 | 13 | 21 | 16 | 32 | 13 | 27 | 15 |
| 8 | 13 | 15 | 6 | 12 | 14 | 22 | 28 | 5 | 26 | 2 | 16 | 1 | 11 | 5 | 23 | 4 | 18 | 11 | 16 |
| 9 | 19 | 12 | 9 | 31 | 7 | 14 | 13 | 23 | 14 | 17 | 22 | 16 | 16 | 14 | 18 | 18 | 32 | 18 | 3 |
| 10 | 3 | 18 | 20 | 24 | 10 | 18 | 5 | 18 | 4 | 12 | 2 | 17 | 9 | 28 | 7 | 6 | 6 | 9 | 17 |
| 11 | 31 | 24 | 28 | 20 | 26 | 28 | 24 | 32 | 31 | 25 | 30 | 32 | 23 | 23 | 27 | 25 | 27 | 22 | 23 |
| 12 | 12 | 20 | 26 | 9 | 1 | 11 | 19 | 11 | 13 | 6 | 3 | 25 | 24 | 20 | 15 | 30 | 8 | 5 | 21 |
| 13 | 22 | 11 | 14 | 27 | 8 | 16 | 16 | 6 | 9 | 3 | 8 | 5 | 4 | 13 | 8 | 14 | 7 | 12 | 2 |
| 14 | 17 | 32 | 29 | 19 | 20 | 24 | 25 | 25 | 18 | 20 | 15 | 28 | 28 | 32 | 26 | 22 | 26 | 19 | 32 |
| 15 | 23 | 30 | 22 | 32 | 4 | 25 | 8 | 30 | 20 | 1 | 29 | 11 | 30 | 15 | 1 | 12 | 22 | 29 | 29 |
| 16 | 26 | 21 | 24 | 23 | 13 | 15 | 27 | 27 | 21 | 7 | 19 | 22 | 20 | 25 | 24 | 17 | 14 | 16 | 28 |
| 17 | 7 | 1 | 2 | 2 | 5 | 12 | 3 | 3 | 5 | 14 | 17 | 15 | 6 | 3 | 28 | 31 | 24 | 20 | 6 |
| 18 | 2 | 5 | 12 | 8 | 11 | 3 | 6 | 4 | 2 | 32 | 9 | 2 | 1 | 9 | 19 | 2 | 25 | 10 | 11 |
| 19 | 6 | 14 | 23 | 13 | 30 | 23 | 23 | 26 | 24 | 31 | 18 | 12 | 21 | 27 | 22 | 26 | 11 | 21 | 25 |
| 20 | 25 | 13 | 16 | 17 | 9 | 32 | 11 | 13 | 15 | 10 | 32 | 7 | 22 | 8 | 13 | 1 | 5 | 28 | 9 |
| 21 | 30 | 31 | 31 | 21 | 25 | 20 | 21 | 31 | 29 | 23 | 20 | 29 | 29 | 31 | 17 | 23 | 31 | 24 | 30 |
| 22 | 14 | 26 | 30 | 18 | 19 | 9 | 26 | 24 | 28 | 22 | 5 | 23 | 15 | 30 | 4 | 27 | 12 | 30 | 26 |
| 23 | 27 | 16 | 25 | 28 | 12 | 21 | 18 | 9 | 23 | 18 | 21 | 19 | 25 | 17 | 14 | 15 | 16 | 26 | 12 |
| 24 | 16 | 4 | 7 | 6 | 6 | 17 | 9 | 1 | 17 | 15 | 14 | 8 | 12 | 1 | 3 | 9 | 21 | 8 | 14 |
| 25 | 20 | 22 | 8 | 1 | 29 | 19 | 10 | 10 | 1 | 13 | 6 | 10 | 3 | 11 | 29 | 13 | 4 | 2 | 7 |
| 26 | 10 | 19 | 4 | 25 | 32 | 29 | 12 | 19 | 3 | 24 | 23 | 21 | 19 | 2 | 2 | 16 | 10 | 14 | 5 |
| 27 | 28 | 27 | 27 | 10 | 22 | 31 | 20 | 28 | 30 | 29 | 4 | 27 | 17 | 29 | 11 | 29 | 20 | 17 | 27 |
| 28 | 29 | 29 | 21 | 5 | 21 | 30 | 22 | 14 | 11 | 4 | 26 | 20 | 10 | 16 | 31 | 21 | 19 | 23 | 22 |
| 29 | 5 | 9 | 19 | 22 | 2 | 2 | 2 | 2 | 25 | 11 | 11 | 13 | 31 | 6 | 6 | 7 | 9 | 4 | 8 |
| 30 | 11 | 25 | 18 | 14 | 3 | 13 | 17 | 15 | 16 | 5 | 10 | 6 | 14 | 4 | 30 | 11 | 1 | 25 | 18 |
| 31 | 32 | 28 | 32 | 30 | 27 | 27 | 31 | 29 | 32 | 26 | 27 | 31 | 32 | 26 | 21 | 28 | 17 | 15 | 31 |
| 32 | 15 | 23 | 15 | 3 | 16 | 26 | 7 | 17 | 10 | 8 | 1 | 9 | 8 | 22 | 25 | 10 | 2 | 1 | 24 |

## Apêndice G – Correlação do *Ranking* com a posição de Classificação das Cultivares

| | Assis (safrinha) | Baixo Grande Ribeiro | Birigui | Brasilândia do Sul | Campo Mourão | Campo Mourão (safrinha) | Cristalina | Dourados (safrinha) | Goianésia | Goiânia | Goiânia (safrinha) | Ipameri | Itumbiara | Londrina | Londrina (safrinha) | Maracaju |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Assis (safrinha)-SP | | 0,2324 | 0,2727 | 0,2412 | 0,4351 | 0,3504 | 0,3372 | 0,2577 | 0,2984 | 0,2284 | 0,0081 | -0,0319 | -0,0554 | 0,2603 | -0,0044 | 0,2991 |
| Baixo Grande Ribeiro | | | 0,3306 | 0,3046 | 0,4289 | 0,0920 | 0,0436 | 0,1210 | 0,3277 | 0,2771 | 0,2188 | 0,1228 | -0,0367 | 0,1957 | -0,2401 | 0,3156 |
| Birigui | | | | 0,2526 | 0,2518 | -0,0048 | 0,4344 | -0,0645 | 0,1419 | 0,0363 | 0,0605 | 0,2276 | -0,1001 | 0,1107 | 0,0433 | 0,1037 |
| Brasilândia do PR | | | | | 0,2969 | -0,3523 | 0,1015 | 0,3244 | 0,1191 | 0,1257 | -0,2320 | 0,0510 | -0,1683 | 0,2694 | 0,1353 | 0,3416 |
| Campo Mourão | | | | | | 0,1554 | 0,4238 | 0,5810 | 0,1767 | 0,5246 | -0,2133 | -0,1488 | 0,0590 | 0,5755 | -0,2632 | 0,2834 |
| Campo Mourão (safrinha) | | | | | | | 0,1331 | 0,1235 | 0,3229 | 0,3589 | 0,4476 | 0,1650 | 0,2493 | 0,0806 | 0,1037 | 0,2155 |
| Cristalina | | | | | | | | 0,1833 | 0,1250 | 0,4443 | 0,2331 | 0,0997 | 0,3120 | 0,3127 | -0,1283 | 0,2243 |
| Dourados (safrinha) | | | | | | | | | 0,1690 | 0,1987 | -0,3779 | 0,0198 | 0,1913 | 0,4018 | -0,2698 | 0,2247 |
| Goianésia | | | | | | | | | | 0,3853 | 0,3207 | 0,3142 | 0,0015 | 0,3215 | -0,2020 | 0,4501 |
| Goiânia | | | | | | | | | | | 0,2724 | -0,0161 | 0,3919 | 0,5671 | -0,0762 | 0,4677 |
| Goiânia (safrinha) | | | | | | | | | | | | 0,2654 | 0,2522 | -0,1738 | 0,0319 | 0,1639 |
| Ipameri | | | | | | | | | | | | | 0,0055 | 0,0367 | 0,1389 | 0,0828 |
| Itumbiara | | | | | | | | | | | | | | -0,1103 | -0,1158 | -0,1191 |
| Londrina | | | | | | | | | | | | | | | -0,2731 | 0,4791 |
| Londrina (safrinha) | | | | | | | | | | | | | | | | -0,0652 |
| Maracaju | | | | | | | | | | | | | | | | |
| Maracaju (safrinha) | | | | | | | | | | | | | | | | |
| Montividiu (safrinha) | | | | | | | | | | | | | | | | |
| Morrinhos | | | | | | | | | | | | | | | | |
| Palmeiras de Goiás | | | | | | | | | | | | | | | | |
| Palotina | | | | | | | | | | | | | | | | |
| Palotina (safrinha) | | | | | | | | | | | | | | | | |
| Patos de Minas | | | | | | | | | | | | | | | | |
| Piracicaba | | | | | | | | | | | | | | | | |
| Planaltina | | | | | | | | | | | | | | | | |
| Ponta Porã | | | | | | | | | | | | | | | | |
| Ponta Porã (safrinha) | | | | | | | | | | | | | | | | |
| Porangatu | | | | | | | | | | | | | | | | |
| Rio Verde | | | | | | | | | | | | | | | | |
| Rio Verde (safrinha) | | | | | | | | | | | | | | | | |
| São Raimundo das Mangabeiras | | | | | | | | | | | | | | | | |
| Sete Lagoas | | | | | | | | | | | | | | | | |
| Sete Lagoas - Cerrado | | | | | | | | | | | | | | | | |
| Sooretama | | | | | | | | | | | | | | | | |
| Uberlândia | | | | | | | | | | | | | | | | |

Apêndice G – Correlação do *Ranking* com a posição de Classificação das Cultivares

| | Maracaju (safrinha) | Montividiu (safrinha) | Morrinhos | Palmeiras de Goiás | Palotina | Palotina (safrinha) | Patos de Minas | Piracicaba | Planaltina | Ponta Porã | Ponta Porã (safrinha) | Porangatu | Rio Verde | Rio Verde (safrinha) | São Raimundo das Mangabeiras |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Assis (safrinha) | 0,3952 | 0,2694 | 0,1158 | 0,1584 | 0,3944 | 0,3383 | 0,3028 | 0,4366 | 0,2779 | 0,1785 | 0,6419 | 0,2287 | 0,3655 | 0,0894 | -0,0253 |
| Baixo Grande Ribeiro | 0,3684 | 0,3603 | 0,2933 | 0,4109 | 0,4014 | 0,0337 | 0,5403 | 0,4003 | 0,2342 | 0,2643 | 0,2775 | 0,7016 | 0,2562 | 0,3893 | -0,0194 |
| Birigui | 0,1536 | -0,0117 | -0,1067 | 0,1672 | 0,3911 | 0,0843 | 0,0770 | 0,3163 | 0,1430 | 0,3482 | 0,3497 | 0,2430 | 0,2830 | -0,0652 | -0,0960 |
| Brasilândia do PR | 0,3240 | -0,0894 | 0,0315 | 0,2412 | 0,1228 | -0,0601 | 0,2049 | 0,2533 | 0,2548 | 0,2771 | 0,4300 | 0,1580 | 0,1294 | 0,0620 | 0,1012 |
| Campo Mourão | 0,3416 | 0,2023 | 0,1609 | 0,2504 | 0,4355 | 0,1074 | 0,5766 | 0,5165 | 0,4315 | 0,4769 | 0,3347 | 0,5315 | 0,1276 | 0,2995 | 0,2533 |
| Campo Mourão (safrinha) | 0,2665 | 0,6404 | 0,4355 | 0,0583 | 0,2826 | 0,5213 | 0,1679 | 0,4930 | 0,2801 | 0,0817 | 0,1015 | 0,2969 | 0,4329 | 0,4567 | -0,0275 |
| Cristalina | 0,2397 | -0,0726 | 0,2071 | 0,1595 | 0,1595 | -0,1602 | 0,3277 | 0,4252 | 0,6261 | 0,3098 | 0,3101 | 0,3094 | 0,4150 | 0,1547 | -0,0312 |
| Dourados (safrinha) | 0,3823 | -0,0143 | 0,0253 | 0,0077 | 0,2009 | 0,1507 | 0,3845 | 0,1136 | 0,2874 | 0,2870 | 0,2166 | 0,2870 | 0,0517 | 0,1598 | 0,1892 |
| Goianésia | 0,5499 | 0,5136 | 0,4329 | 0,4850 | 0,3394 | 0,2918 | 0,5546 | 0,4879 | 0,3930 | 0,1177 | 0,4201 | 0,4648 | 0,4069 | 0,3893 | -0,0143 |
| Goiânia | 0,3735 | 0,4923 | 0,6265 | 0,4809 | 0,2020 | 0,1972 | 0,6375 | 0,6441 | 0,6782 | 0,3160 | 0,0810 | 0,5385 | 0,3897 | 0,6646 | 0,0326 |
| Goiânia (safrinha) | 0,2177 | 0,4062 | 0,4223 | 0,0139 | 0,3145 | 0,0422 | 0,2797 | 0,3394 | 0,2207 | -0,0224 | 0,0011 | 0,2368 | 0,2988 | 0,3970 | 0,0444 |
| Ipameri | 0,3310 | 0,2386 | 0,3732 | 0,1408 | 0,0550 | 0,3013 | 0,0026 | 0,1994 | 0,2159 | 0,0609 | 0,1712 | 0,2240 | 0,5205 | -0,0147 | -0,0598 |
| Itumbiara | 0,3713 | 0,1063 | 0,1551 | 0,2529 | -0,1833 | 0,1345 | 0,1998 | 0,0048 | 0,3933 | -0,2375 | 0,0898 | 0,0751 | 0,2830 | 0,1169 | -0,0598 |
| Londrina | 0,0795 | 0,1389 | 0,2889 | 0,2067 | 0,2841 | 0,1664 | 0,3765 | 0,5685 | 0,4582 | 0,3908 | 0,0341 | 0,4472 | 0,1081 | 0,4069 | 0,1037 |
| Londrina (safrinha) | -0,2049 | 0,0268 | -0,0055 | -0,2086 | -0,0678 | 0,0183 | -0,4267 | -0,0737 | -0,0139 | 0,0612 | -0,0839 | -0,2749 | 0,1826 | -0,1081 | -0,1466 |
| Maracaju | 0,2749 | 0,2042 | 0,4091 | 0,2247 | 0,4131 | 0,0297 | 0,5506 | 0,4146 | 0,4923 | 0,5323 | 0,2625 | 0,3985 | 0,3218 | 0,5960 | -0,0480 |
| Maracaju (safrinha) | | 0,5678 | 0,5062 | 0,3614 | 0,2053 | 0,5312 | 0,5623 | 0,4732 | 0,5304 | 0,0022 | 0,5022 | 0,4534 | 0,4256 | 0,3266 | 0,1188 |
| Montividiu (safrinha) | | | 0,7265 | 0,2504 | 0,1540 | 0,6158 | 0,4245 | 0,6089 | 0,3798 | -0,1045 | 0,0209 | 0,4556 | 0,3823 | 0,5447 | 0,1752 |
| Morrinhos | | | | 0,3633 | 0,0213 | 0,3911 | 0,4516 | 0,6199 | 0,6485 | 0,1239 | -0,0337 | 0,6195 | 0,5352 | 0,7309 | 0,0612 |
| Palmeiras de Goiás | | | | | -0,0169 | 0,0473 | 0,3464 | 0,3959 | 0,3530 | 0,1298 | 0,3981 | 0,3878 | 0,5227 | 0,2793 | -0,4307 |
| Palotina | | | | | | 0,1701 | 0,3739 | 0,3977 | 0,1059 | 0,5729 | 0,2324 | 0,3442 | 0,0902 | 0,3908 | 0,0631 |
| Palotina (safrinha) | | | | | | | 0,1239 | 0,4025 | 0,2236 | -0,1488 | 0,1734 | 0,1697 | 0,2133 | 0,1888 | 0,1664 |
| Patos de Minas | | | | | | | | 0,4421 | 0,6250 | 0,2592 | 0,3570 | 0,5337 | 0,2573 | 0,4622 | 0,2430 |
| Piracicaba | | | | | | | | | 0,5106 | 0,2962 | 0,2900 | 0,6140 | 0,4604 | 0,6551 | 0,0942 |
| Planaltina | | | | | | | | | | 0,2133 | 0,2845 | 0,4674 | 0,5744 | 0,4274 | 0,0319 |
| Ponta Porã | | | | | | | | | | | 0,1569 | 0,4729 | 0,1620 | 0,3376 | -0,1133 |
| Ponta Porã (safrinha) | | | | | | | | | | | | 0,1778 | 0,4560 | -0,1422 | 0,0436 |
| Porangatu | | | | | | | | | | | | | 0,4402 | 0,6393 | 0,0253 |
| Rio Verde | | | | | | | | | | | | | | 0,3134 | -0,2892 |
| Rio Verde (safrinha) | | | | | | | | | | | | | | | 0,0110 |
| São Raimundo das Mangabeiras | | | | | | | | | | | | | | | |
| Sete Lagoas | | | | | | | | | | | | | | | |
| Sete Lagoas - Cerrado | | | | | | | | | | | | | | | |
| Sooretama | | | | | | | | | | | | | | | |
| Uberlândia | | | | | | | | | | | | | | | |

Apêndice G – Correlação do *Ranking* com a posição de Classificação das Cultivares

| | Sete Lagoas | Sete Lagoas - Cerrado | Sooretama | Uberlândia |
|---|---|---|---|---|
| Assis (safrinha) | -0,0788 | 0,4751 | 0,1870 | 0,3460 |
| Baixo Grande Ribeiro | 0,3783 | 0,2958 | 0,2962 | 0,3189 |
| Birigui | 0,0598 | 0,3138 | -0,0055 | -0,0920 |
| Brasilândia do PR | 0,0803 | 0,4098 | 0,2841 | -0,0656 |
| Campo Mourão | 0,1789 | 0,4392 | 0,2159 | 0,3416 |
| Campo Mourão (safrinha) | 0,3435 | -0,0326 | 0,2174 | 0,6228 |
| Cristalina | 0,1562 | 0,4520 | 0,0993 | 0,2874 |
| Dourados (safrinha) | 0,2394 | 0,2405 | 0,0693 | 0,1514 |
| Goianésia | 0,2724 | 0,1778 | 0,5601 | 0,4362 |
| Goiânia | 0,3581 | 0,3721 | 0,4161 | 0,7665 |
| Goiânia (safrinha) | 0,3501 | 0,0033 | 0,2258 | 0,4523 |
| Ipameri | 0,0557 | -0,2408 | 0,0788 | 0,1749 |
| Itumbiara | 0,1789 | 0,2199 | 0,1606 | 0,2555 |
| Londrina | 0,0942 | 0,2427 | 0,1565 | 0,3871 |
| Londrina (safrinha) | -0,0187 | -0,1569 | -0,0975 | 0,0048 |
| Maracaju | 0,1734 | 0,3032 | 0,4993 | 0,4883 |
| Maracaju (safrinha) | 0,2207 | 0,3325 | 0,4131 | 0,4476 |
| Montividiu (safrinha) | 0,2933 | -0,0231 | 0,2951 | 0,7566 |
| Morrinhos | 0,3761 | 0,0484 | 0,3457 | 0,7914 |
| Palmeiras de Goiás | 0,0495 | 0,2632 | 0,5018 | 0,2027 |
| Palotina | 0,3240 | 0,1323 | 0,1089 | 0,1716 |
| Palotina (safrinha) | 0,1279 | -0,1463 | 0,1261 | 0,3717 |
| Patos de Minas | 0,3171 | 0,3900 | 0,4663 | 0,5733 |
| Piracicaba | 0,3094 | 0,2933 | 0,3427 | 0,6455 |
| Planaltina | 0,2848 | 0,3149 | 0,3757 | 0,5913 |
| Ponta Porã | 0,3233 | 0,4501 | 0,2768 | 0,1521 |
| Ponta Porã (safrinha) | -0,0598 | 0,5315 | 0,4894 | 0,0436 |
| Porangatu | 0,6221 | 0,4062 | 0,4091 | 0,5726 |
| Rio Verde | 0,2529 | 0,1877 | 0,3420 | 0,4432 |
| Rio Verde (safrinha) | 0,5440 | 0,1265 | 0,2889 | 0,7199 |
| São Raimundo das Mangabeiras | 0,0506 | 0,0645 | -0,1287 | 0,1327 |
| Sete Lagoas | | 0,1202 | 0,3519 | 0,3856 |
| Sete Lagoas - Cerrado | | | 0,3625 | 0,2045 |
| Sooretama | | | | 0,3864 |
| Uberlândia | | | | |

# Apêndice H – *Ranking* da Correlação de Spearman entre Experimentos

| Posição | Cidade 1 | Cidade 2 | Correlação Spearman | Posição | Cidade 1 | Cidade 2 | Correlação Spearman |
|---|---|---|---|---|---|---|---|
| 1 | Morrinhos | Uberlândia | 0,7914 | 55 | Maracaju (safrinha) | Planaltina | 0,5304 |
| 2 | Goiânia | Uberlândia | 0,7665 | 56 | Campo Mourão | Goiânia | 0,5246 |
| 3 | Montividiu (safrinha) | Uberlândia | 0,7566 | 57 | Palmeiras de Goiás | Rio Verde | 0,5227 |
| 4 | Morrinhos | Rio Verde (safrinha) | 0,7309 | 58 | Campo Mourão (safrinha) | Palotina (safrinha) | 0,5213 |
| 5 | Montividiu (safrinha) | Morrinhos | 0,7265 | 59 | Ipameri | Rio Verde | 0,5205 |
| 6 | Rio Verde (safrinha) | Uberlândia | 0,7199 | 60 | Campo Mourão | Piracicaba | 0,5165 |
| 7 | Baixo Grande Ribeiro | Porangatu | 0,7016 | 61 | Goianésia | Montividiu (safrinha) | 0,5136 |
| 8 | Goiânia | Planaltina | 0,6782 | 62 | Piracicaba | Planaltina | 0,5106 |
| 9 | Goiânia | Rio Verde (safrinha) | 0,6646 | 63 | Maracaju (safrinha) | Morrinhos | 0,5062 |
| 10 | Piracicaba | Rio Verde (safrinha) | 0,6551 | 64 | Maracaju (safrinha) | Ponta Porã (safrinha) | 0,5022 |
| 11 | Morrinhos | Planaltina | 0,6485 | 65 | Palmeiras de Goiás | Sooretama | 0,5018 |
| 12 | Piracicaba | Uberlândia | 0,6455 | 66 | Maracaju | Sooretama | 0,4993 |
| 13 | Goiânia | Piracicaba | 0,6441 | 67 | Campo Mourão (safrinha) | Piracicaba | 0,4930 |
| 14 | Assis | Ponta Porã (safrinha) | 0,6419 | 68 | Goiânia | Montividiu (safrinha) | 0,4923 |
| 15 | Campo Mourão (safrinha) | Montividiu (safrinha) | 0,6404 | 69 | Maracaju | Planaltina | 0,4923 |
| 16 | Porangatu | Rio Verde (safrinha) | 0,6393 | 70 | Ponta Porã (safrinha) | Sooretama | 0,4894 |
| 17 | Goiânia | Patos de Minas | 0,6375 | 71 | Maracaju | Uberlândia | 0,4883 |
| 18 | Goiânia | Morrinhos | 0,6265 | 72 | Goianésia | Piracicaba | 0,4879 |
| 19 | Cristalina | Planaltina | 0,6261 | 73 | Goianésia | Palmeiras de Goiás | 0,4850 |
| 20 | Patos de Minas | Planaltina | 0,6250 | 74 | Goiânia | Palmeiras de Goiás | 0,4809 |
| 21 | Campo Mourão (safrinha) | Uberlândia | 0,6228 | 75 | Londrina | Maracaju | 0,4791 |
| 22 | Porangatu | Sete Lagoas | 0,6221 | 76 | Campo Mourão | Ponta Porã | 0,4769 |
| 23 | Morrinhos | Piracicaba | 0,6199 | 77 | Assis | Sete Lagoas - Cerrado | 0,4751 |
| 24 | Morrinhos | Porangatu | 0,6195 | 78 | Maracaju (safrinha) | Piracicaba | 0,4732 |
| 25 | Montividiu (safrinha) | Palotina (safrinha) | 0,6158 | 79 | Ponta Porã | Porangatu | 0,4729 |
| 26 | Piracicaba | Porangatu | 0,6140 | 80 | Goiânia | Maracaju | 0,4677 |
| 27 | Montividiu (safrinha) | Piracicaba | 0,6089 | 81 | Planaltina | Porangatu | 0,4674 |
| 28 | Maracaju | Rio Verde (safrinha) | 0,5960 | 82 | Patos de Minas | Sooretama | 0,4663 |
| 29 | Planaltina | Uberlândia | 0,5913 | 83 | Goianésia | Porangatu | 0,4648 |
| 30 | Campo Mourão | Dourados (safrinha) | 0,5810 | 84 | Patos de Minas | Rio Verde (safrinha) | 0,4622 |
| 31 | Campo Mourão | Patos de Minas | 0,5766 | 85 | Piracicaba | Rio Verde | 0,4604 |
| 32 | Campo Mourão | Londrina | 0,5755 | 86 | Londrina | Planaltina | 0,4582 |
| 33 | Planaltina | Rio Verde | 0,5744 | 87 | Campo Mourão (safrinha) | Rio Verde (safrinha) | 0,4567 |
| 34 | Patos de Minas | Uberlândia | 0,5733 | 88 | Ponta Porã (safrinha) | Rio Verde | 0,4560 |
| 35 | Palotina | Ponta Porã | 0,5729 | 89 | Montividiu (safrinha) | Porangatu | 0,4556 |
| 36 | Porangatu | Uberlândia | 0,5726 | 90 | Maracaju (safrinha) | Porangatu | 0,4534 |
| 37 | Londrina | Piracicaba | 0,5685 | 91 | Goiânia (safrinha) | Uberlândia | 0,4523 |
| 38 | Maracaju (safrinha) | Montividiu (safrinha) | 0,5678 | 92 | Cristalina | Sete Lagoas - Cerrado | 0,4520 |
| 39 | Goiânia | Londrina | 0,5671 | 93 | Morrinhos | Patos de Minas | 0,4516 |
| 40 | Maracaju (safrinha) | Patos de Minas | 0,5623 | 94 | Goianésia | Maracaju | 0,4501 |
| 41 | Goianésia | Sooretama | 0,5601 | 95 | Ponta Porã | Sete Lagoas - Cerrado | 0,4501 |
| 42 | Goianésia | Patos de Minas | 0,5546 | 96 | Campo Mourão (safrinha) | Goiânia (safrinha) | 0,4476 |
| 43 | Maracaju | Patos de Minas | 0,5506 | 97 | Maracaju (safrinha) | Uberlândia | 0,4476 |
| 44 | Goianésia | Maracaju (safrinha) | 0,5499 | 98 | Londrina | Porangatu | 0,4472 |
| 45 | Montividiu (safrinha) | Rio Verde (safrinha) | 0,5447 | 99 | Cristalina | Goiânia | 0,4443 |
| 46 | Rio Verde (safrinha) | Sete Lagoas | 0,5440 | 100 | Rio Verde | Uberlândia | 0,4432 |
| 47 | Baixo Grande Ribeiro | Patos de Minas | 0,5403 | 101 | Patos de Minas | Piracicaba | 0,4421 |
| 48 | Goiânia | Porangatu | 0,5385 | 102 | Porangatu | Rio Verde | 0,4402 |
| 49 | Morrinhos | Rio Verde | 0,5352 | 103 | Campo Mourão | Sete Lagoas - Cerrado | 0,4392 |
| 50 | Patos de Minas | Porangatu | 0,5337 | 104 | Assis | Piracicaba | 0,4366 |
| 51 | Maracaju | Ponta Porã | 0,5323 | 105 | Goianésia | Uberlândia | 0,4362 |
| 52 | Campo Mourão | Porangatu | 0,5315 | 106 | Campo Mourão | Palotina | 0,4355 |
| 53 | Ponta Porã (safrinha) | Sete Lagoas - Cerrado | 0,5315 | 107 | Campo Mourão (safrinha) | Morrinhos | 0,4355 |
| 54 | Maracaju (safrinha) | Palotina (safrinha) | 0,5312 | 108 | Assis | Campo Mourão | 0,4351 |

# Apêndice H – *Ranking* da Correlação de Spearman entre Experimentos

| Posição | Cidade 1 | Cidade 2 | Correlação Spearman | Posição | Cidade 1 | Cidade 2 | Correlação Spearman |
|---|---|---|---|---|---|---|---|
| 109 | Birigui | Cristalina | 0,4344 | 163 | Dourados (safrinha) | Maracaju (safrinha) | 0,3823 |
| 110 | Campo Mourão (safrinha) | Rio Verde | 0,4329 | 164 | Montividiu (safrinha) | Rio Verde | 0,3823 |
| 111 | Goianésia | Morrinhos | 0,4329 | 165 | Montividiu (safrinha) | Planaltina | 0,3798 |
| 112 | Campo Mourão | Planaltina | 0,4315 | 166 | Baixo Grande Ribeiro | Sete Lagoas | 0,3783 |
| 113 | Brasilândia do Sul | Ponta Porã (safrinha) | 0,4300 | 167 | Londrina | Patos de Minas | 0,3765 |
| 114 | Baixo Grande Ribeiro | Campo Mourão | 0,4289 | 168 | Morrinhos | Sete Lagoas | 0,3761 |
| 115 | Planaltina | Rio Verde (safrinha) | 0,4274 | 169 | Planaltina | Sooretama | 0,3757 |
| 116 | Maracaju (safrinha) | Rio Verde | 0,4256 | 170 | Palotina | Patos de Minas | 0,3739 |
| 117 | Cristalina | Piracicaba | 0,4252 | 171 | Goiânia | Maracaju (safrinha) | 0,3735 |
| 118 | Montividiu (safrinha) | Patos de Minas | 0,4245 | 172 | Ipameri | Morrinhos | 0,3732 |
| 119 | Campo Mourão | Cristalina | 0,4238 | 173 | Goiânia | Sete Lagoas - Cerrado | 0,3721 |
| 120 | Goiânia (safrinha) | Morrinhos | 0,4223 | 174 | Palotina (safrinha) | Uberlândia | 0,3717 |
| 121 | Goianésia | Ponta Porã (safrinha) | 0,4201 | 175 | Itumbiara | Maracaju (safrinha) | 0,3713 |
| 122 | Goiânia | Sooretama | 0,4161 | 176 | Baixo Grande Ribeiro | Maracaju (safrinha) | 0,3684 |
| 123 | Cristalina | Rio Verde | 0,4150 | 177 | Assis | Rio Verde | 0,3655 |
| 124 | Maracaju | Piracicaba | 0,4146 | 178 | Morrinhos | Palmeiras de Goiás | 0,3633 |
| 125 | Maracaju | Palotina | 0,4131 | 179 | Sete Lagoas - Cerrado | Sooretama | 0,3625 |
| 126 | Maracaju (safrinha) | Sooretama | 0,4131 | 180 | Maracaju (safrinha) | Palmeiras de Goiás | 0,3614 |
| 127 | Baixo Grande Ribeiro | Palmeiras de Goiás | 0,4109 | 181 | Baixo Grande Ribeiro | Montividiu (safrinha) | 0,3603 |
| 128 | Brasilândia do Sul | Sete Lagoas - Cerrado | 0,4098 | 182 | Campo Mourão (safrinha) | Goiânia | 0,3589 |
| 129 | Maracaju | Morrinhos | 0,4091 | 183 | Goiânia | Sete Lagoas | 0,3581 |
| 130 | Porangatu | Sooretama | 0,4091 | 184 | Patos de Minas | Ponta Porã (safrinha) | 0,3570 |
| 131 | Goianésia | Rio Verde | 0,4069 | 185 | Palmeiras de Goiás | Planaltina | 0,3530 |
| 132 | Londrina | Rio Verde (safrinha) | 0,4069 | 186 | Sete Lagoas | Sooretama | 0,3519 |
| 133 | Goiânia (safrinha) | Montividiu (safrinha) | 0,4062 | 187 | Assis | Campo Mourão (safrinha) | 0,3504 |
| 134 | Porangatu | Sete Lagoas - Cerrado | 0,4062 | 188 | Goiânia (safrinha) | Sete Lagoas | 0,3501 |
| 135 | Palotina (safrinha) | Piracicaba | 0,4025 | 189 | Birigui | Ponta Porã (safrinha) | 0,3497 |
| 136 | Dourados (safrinha) | Londrina | 0,4018 | 190 | Birigui | Ponta Porã | 0,3482 |
| 137 | Baixo Grande Ribeiro | Palotina | 0,4014 | 191 | Palmeiras de Goiás | Patos de Minas | 0,3464 |
| 138 | Baixo Grande Ribeiro | Piracicaba | 0,4003 | 192 | Assis | Uberlândia | 0,3460 |
| 139 | Maracaju | Porangatu | 0,3985 | 193 | Morrinhos | Sooretama | 0,3457 |
| 140 | Palmeiras de Goiás | Ponta Porã (safrinha) | 0,3981 | 194 | Palotina | Porangatu | 0,3442 |
| 141 | Palotina | Piracicaba | 0,3977 | 195 | Campo Mourão (safrinha) | Sete Lagoas | 0,3435 |
| 142 | Goiânia (safrinha) | Rio Verde (safrinha) | 0,3970 | 196 | Piracicaba | Sooretama | 0,3427 |
| 143 | Palmeiras de Goiás | Piracicaba | 0,3959 | 197 | Rio Verde | Sooretama | 0,3420 |
| 144 | Assis | Maracaju (safrinha) | 0,3952 | 198 | Brasilândia do Sul | Maracaju | 0,3416 |
| 145 | Assis | Palotina | 0,3944 | 199 | Campo Mourão | Maracaju (safrinha) | 0,3416 |
| 146 | Itumbiara | Planaltina | 0,3933 | 200 | Campo Mourão | Uberlândia | 0,3416 |
| 147 | Goianésia | Planaltina | 0,3930 | 201 | Goianésia | Palotina | 0,3394 |
| 148 | Goiânia | Itumbiara | 0,3919 | 202 | Goiânia (safrinha) | Piracicaba | 0,3394 |
| 149 | Birigui | Palotina | 0,3911 | 203 | Assis | Palotina (safrinha) | 0,3383 |
| 150 | Morrinhos | Palotina (safrinha) | 0,3911 | 204 | Ponta Porã | Rio Verde (safrinha) | 0,3376 |
| 151 | Londrina | Ponta Porã | 0,3908 | 205 | Assis | Cristalina | 0,3372 |
| 152 | Palotina | Rio Verde (safrinha) | 0,3908 | 206 | Campo Mourão | Ponta Porã (safrinha) | 0,3347 |
| 153 | Patos de Minas | Sete Lagoas - Cerrado | 0,3900 | 207 | Maracaju (safrinha) | Sete Lagoas - Cerrado | 0,3325 |
| 154 | Goiânia | Rio Verde | 0,3897 | 208 | Ipameri | Maracaju (safrinha) | 0,3310 |
| 155 | Baixo Grande Ribeiro | Rio Verde (safrinha) | 0,3893 | 209 | Baixo Grande Ribeiro | Birigui | 0,3306 |
| 156 | Goianésia | Rio Verde (safrinha) | 0,3893 | 210 | Baixo Grande Ribeiro | Goianésia | 0,3277 |
| 157 | Palmeiras de Goiás | Porangatu | 0,3878 | 211 | Cristalina | Patos de Minas | 0,3277 |
| 158 | Londrina | Uberlândia | 0,3871 | 212 | Maracaju (safrinha) | Rio Verde (safrinha) | 0,3266 |
| 159 | Sooretama | Uberlândia | 0,3864 | 213 | Brasilândia do Sul | Dourados (safrinha) | 0,3244 |
| 160 | Sete Lagoas | Uberlândia | 0,3856 | 214 | Brasilândia do Sul | Maracaju (safrinha) | 0,3240 |
| 161 | Goianésia | Goiânia | 0,3853 | 215 | Palotina | Sete Lagoas | 0,3240 |
| 162 | Dourados (safrinha) | Patos de Minas | 0,3845 | 216 | Ponta Porã | Sete Lagoas | 0,3233 |