UNIVERSIDADE ESTADUAL DE PONTA GROSSA
SETOR DE CIÊNCIAS AGRÁRIAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

RICARDO KAZUO BABA

CORREÇÃO DE DADOS AGROMETEOROLÓGICOS UTILIZANDO MÉTODOS ESTATÍSTICOS

PONTA GROSSA
2012

RICARDO KAZUO BABA

CORREÇÃO DE DADOS AGROMETEOROLÓGICOS UTILIZANDO MÉTODOS ESTATÍSTICOS

Dissertação submetida ao Programa de Pós-Graduação em Computação Aplicada, da Universidade Estadual de Ponta Grossa, como requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Orientadora: Profa. Dra. Maria Salete M. Gomes Vaz

PONTA GROSSA
2012

Ficha Catalográfica Elaborada pelo Setor Tratamento da Informação Bicen / UEPG

Baba, Ricardo Kazuo
B112c Correção de dados agrometeorológicos utilizando métodos estatísticos
Ponta Grossa, 2012.
89 f.

Dissertação (Mestrado em Computação Aplicada), Universidade Estadual de Ponta Grossa.
Orientadora: Profa. Dra.Maria Salete M. Gomes Vaz.

1. Dados climáticos – agricultura. 2. Banco de dados espacial. 3. Dados meteorológicos. I. Vaz, Maria Salete M. Gomes. II. Universidade Estadual de Ponta Grossa. Mestrado em Computação Aplicada. III. T.

CDD: 005.7

TERMO DE APROVAÇÃO

RICARDO KAZUO BABA

"CORREÇÃO DE DADOS AGROMETEOROLÓGICOS UTILIZANDO MÉTODOS ESTATÍSTICOS"

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora.

Orientador: Maria Salete Marcon Gomes Vaz
UEPG

Marcos Sfair Sunye
UFPR

Selma Regina Aranha Ribeiro
UEPG

Ponta Grossa, 31 de Julho de 2012.

AGRADECIMENTOS

A Deus pelo dom da vida e por sempre me dar forças para continuar.

A minha esposa Kélly e aos meus filhos Amanda e Rafael pela paciência, apoio, incentivo e compreensão.

A toda a minha família, meus pais e irmãos pela ajuda e apoio em todos os momentos ao longo do mestrado.

A todos os professores e colegas do programa de Mestrado em Computação Aplicada à Agricultura por compartilharem de seus conhecimentos.

A Fundação ABC e seus colaboradores por disponibilizar o material necessário para esta pesquisa.

A minha orientadora Profa. Dra. Maria Salete Marcon Gomes Vaz, pela sua ajuda orientação e dedicação para comigo.

RESUMO

A análise de dados climáticos serve de suporte na previsão de fenômenos relacionados, na avaliação de seus dados históricos e para a tomada de decisões, em especial na área da agricultura. Garantir a sua qualidade é fundamental. O processo de coleta desses dados, através das estações meteorológicas, pode apresentar problemas, onde dados inconsistentes podem ser geridos ou obtidos. A identificação de dados inconsistentes ou suspeitos é de fundamental importância na garantia de qualidade dos dados. Este trabalho apresenta uma abordagem para solução do problema, utilizando técnicas estatísticas e geoestatísticas na identificação de dados inconsistentes e na estimativa de dados a serem corrigidos ou preenchidos. A implementação destas técnicas em um banco de dados espacial apresentou-se como um facilitador na identificação e no preenchimento desses dados. Para avaliação destas técnicas utilizou-se de dados das estações localizadas no Estado do Paraná, para análise da variável temperatura. Para avaliar os resultados, foram utilizados os erros médio e quadrático. Como resultado, destaca-se que as técnicas de identificação de erros mostraram-se adequadas na consistência de erros básicos e históricos. A validação espacial apresentou baixo desempenho por superestimar a quantidade de dados incorretos. Quanto as técnicas utilizadas na estimativa dos dados, Krigagem, Inverso da Distância e Regressão Linear, todas apresentaram desempenho semelhantes com relação à análise dos erros.

Palavras-chave: preenchimento de falhas; dados meteorológicos; estatística; geoestatística; controle de qualidade de dados; banco de dados espacial.

ABSTRACT

Climatic data are more and more important to predict climate phenomena or to evaluate historical data that serve as support for decision making especially for agriculture. Ensuring the quality of these data is crucial. These data are collected by the meteorological stations, during this process some data gaps and data inconsistent may be generated. Identify suspicious or inconsistent data is very important to ensure data quality. This paper presents an approach that uses statistical and geostatistical techniques to identify incorrect and suspicious data and estimate new values to fill gaps and errors. In this research, a spatial database was used to implement these techniques (statistical and geostatistical) and to test and evaluate the weather data. To evaluate these techniques we used data from stations located in Paraná State to evaluate the temperature variable. To check the results of the estimated data, we used the mean absolute error (MAE) and the root mean square error (RMSE). As a result, the uses of these techniques have proved to be suitable to identify basic errors and historical errors. The temporal validation showed a poor performance by overestimating the amount of incorrect data. Regarding the estimation techniques applied Kriging, Inverse of Distance Weighted and Linear Regression, all showed similar performance in the error analysis.

Keywords: gap filling; meteorological data; statistics; geostatistics; data quality control; spatial database.

LISTA DE SIGLAS

EM – Erro Médio

EMA – Erro Médio Absoluto

EMP – Erro Médio Percentual

IQD – Inverso do quadrado da distância

REQM – Raiz Erro Quadrático Médio

R.L. – Regressão Linear

SGBD – Sistema gerenciador de banco de dados

SGBDOR – Sistema gerenciador de banco de dados objeto-relacional

SIG – Sistema de Informação geográfica.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Para toda e qualquer instituição, seja ela privada, pública, acadêmica, industrial, comercial ou governamental, os dados representam as informações em suas áreas de atuação. De posse dessas informações, considerados como patrimônio das organizações, é possível analisar e tomar decisões (VENTURA, 2012).

Sob esse ponto de vista, pode-se afirmar que os dados climáticos são vitais na agricultura, pois servem de base para aperfeiçoar processos de manejo ou mesmo para verificar a viabilidade do plantio de determinadas culturas. Estes dados podem ser utilizados para dar suporte na predição de doenças e pragas relacionadas a algumas culturas, podendo servir de base para sistemas de alertas de doenças (TRENTIN et. al., 2009).

Os dados climáticos, geralmente, são obtidos de forma automática por meio de estações meteorológicas, as quais coletam dados de vários sensores, 24 horas por dia, em intervalos de tempo variados, gerando assim uma grande quantidade de dados e/ou informação.

Quando a leitura dos dados climáticos não é realizada de maneira automática, é necessário que um observador anote as medidas captadas pelos equipamentos. No entanto, problemas diversos podem ocorrer e impedir que o observador realize alguma leitura dos equipamentos, em determinado horário do dia, provocando falha na leitura dos dados (VENTURA, 2012).

Uma vez de posse destes dados, os mesmos devem ser analisados, e caso sejam identificadas falhas ou erros, essas devem ser corrigidas da forma mais rápida e consistente possível.

Segundo Tsukahara et al. (2010), a ausência de registros meteorológicos é um problema frequente na maioria das séries climatológicas brasileiras. Existem vários fatores que podem influenciar na ausência ou no erro de dados de uma estação meteorológica, dentre elas pode-se citar falhas nos sensores, calibração dos equipamentos, falhas nas transmissões dos dados (telemetria), manutenção nos

sistema e intervenção de agentes externos.

Segundo Ventura (2012), os fatores que podem influenciar na coleta dos dados são (entre outros): ações de animais, próximos aos equipamentos; e as fortes mudanças no tempo. Deve-se, ainda, levar em consideração o fato de que os sensores são aparelhos eletrônicos e estão sujeitos a falhas. Todos esses fatores podem resultar em erros de leitura ou mesmo em ausência de dados, as quais precisam ser tratadas e devidamente corrigidas.

Com base nesse cenário, as coletas de grande quantidade de dados, podem ocorrer de forma automática ou manual, por longos períodos de tempo e com resoluções variadas. Muitas destas coletas podem ocorrer em intervalos de tempo bastante curtos, como por exemplo, em estações que realizam coletas de dados em intervalos de quinze em quinze minutos.

As ausências ou inconsistências podem gerar dificuldades, dúvidas ou imprecisão na análise dos dados. Desta forma, faz parte desta pesquisa a aplicação de métodos estatísticos e geoestatísticos em conjunto com recursos computacionais, como os bancos de dados espaciais, que visem auxiliar na identificação de dados inconsistentes ou suspeitos e, também, permitir a sugestão de valores, tanto para os dados suspeitos, quanto para as séries faltantes.

1.2 JUSTIFICATIVA

Devido ao grande volume de dados gerados pelas estações meteorológicas, e devido aos mais diversos tipos de problemas, tanto na coleta quanto na transmissão, faz-se necessária uma análise destes dados, a fim de identificar dados inconsistentes ou nulos. Esta avaliação tem como principal finalidade evitar que dados inconsistentes ou falhos sejam utilizados como informações reais, dos fenômenos meteorológicos, os quais representam.

No processo de análise dos dados deve-se definir qual o método mais adequado para tratamento daqueles considerados inconsistentes ou nulos. A respeito destes problemas, deve-se decidir se estes dados serão descartados ou se serão estimados.

Um método bastante comum baseia-se na utilização da média das medições mais próximas para estimar o valor faltante ou inconsistente. Outro método baseia-

se na repetição de valores para preenchimento da série faltante. Entretanto, esses são métodos que podem apresentar características como: a série de dados já não irá representar completamente a realidade e, também, há possibilidade de erros nas séries utilizadas para o cálculo da média ou repetição, já que são necessárias leituras de períodos próximos, para preencher a falha. Este problema pode ser ainda mais evidente quando as séries de dados mais próximas representem princípios de falhas nos equipamentos de coleta.

Com o intuito de estimar valores mais próximos do valor real, uma das alternativas é a automação mediante sistemas computacionais apoiado em estatística para auxiliar na identificação dos erros e na estimativa de valores no preenchimento das falhas.

Atualmente, vários métodos estão sendo utilizados e estudados com o intuito de obter maior acurácia nas estimativas de dados. Dentre eles podemos citar os métodos baseados em estatística e geoestatística (BOTELHO et al, 2005; NOGUEIRA & AMARAL, 2009, SOUZA et al, 2011; VIEIRA, 2000; VIOLA et al. , 2010) e, também, os métodos baseados em Inteligência Artificial (TSUKAHARA et al, 2010; VENTURA, 2012).

Os métodos baseados em estatística visam estimar os dados faltantes por meio da análise de séries históricas e de métodos de interpolação. Os métodos estatísticos podem ser utilizados para verificar a correlação entre os dados, assim como se um determinado dado está dentro de um intervalo de confiança, com base em seu comportamento na série histórica.

A utilização combinada de técnicas e métodos tem como meta facilitar, auxiliar e automatizar a identificação de dados inconsistentes, assim como estimar os seus valores.

A automatização do processo traz benefícios, especialmente, devido ao volume de dados gerados pelas estações e que dificilmente pode ser analisado com precisão, quando feito manualmente. Esse processo automático permite uma avaliação criteriosa e consistente dos dados de entrada, melhorando a qualidade das informações que serão utilizadas em análises posteriores.

1.3 OBJETIVOS

O objetivo geral desta pesquisa é elaborar uma metodologia para automatizar e dar suporte na identificação e estimativa dos dados meteorológicos inconsistentes em um banco de dados espacial.

Os objetivos específicos são os que seguem.

- Implementar técnicas estatísticas e geoestatísticas de validação e estimativa de dados climáticos em um banco de dados espacial;
- Testar o método de Mateo & Leung (2010) de validação de dados climáticos;
- Analisar os modelos estatísticos associados ao banco de dados espacial no que tange a consistência dos dados climáticos;
- Avaliar os resultados de cada técnica aplicada quanto a sua precisão na estimativa dos dados; e
- Verificar a precisão de cada técnica na estimativa de dados para a região de estudo;

1.4 ESTRUTURA DO TRABALHO

Este trabalho está estruturado em 5 (cinco) capítulos, incluindo este capítulo de introdução. No Capítulo 2, Fundamentação Teórica, são descritos os conceitos básicos inerentes a estatística, geoestatística, bancos de dados espaciais e de dados meteorológicos. No Capítulo 3, Materiais e Métodos, é apresentada a metodologia utilizada ao longo da pesquisa. No Capítulo 4, Resultados e Discussão, é apresentado os resultados obtidos, além de descrever trabalhos relacionados. E, finalmente, no Capítulo 5 são realizadas as conclusões e apresentadas as perspectivas de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A aplicação de métodos estatísticos é amplamente utilizada nas mais variadas área de conhecimento, sejam para estimar valores, relações ou para descrição de resultados.

De acordo com Guimarães (2004), os métodos clássicos de análise estatística de dados, geralmente, supõem que as realizações das variáveis aleatórias são independentes entre si, ou seja, que observações vizinhas não exercem influências umas sobre as outras. No entanto, alguns fenômenos da natureza apresentam certa estruturação, de acordo com as variações da vizinhança, portanto neste caso as variações não são aleatórias e apresentam algum grau de dependência espacial.

A análise espacial de dados apresenta-se como uma alternativa ou complemento da análise clássica de dados, sendo que considera as correlações entre as observações quando se realiza estimativas. Alguns métodos de análise espacial têm sido estudados, dentre eles o geoestatístico.

Neste capítulo serão abordados conceitos básicos tanto de estatística quanto de geoestatística apresentando uma visão geral de seus principais componentes e métodos. Na sequencia são apresentados os conceitos básicos de dados espaciais e dados meteorológicos.

2.1 ESTATÍSTICA

Estatística é a ciência que investiga os processos de obtenção, organização e análise de dados sobre uma população, e os métodos para tirar conclusões ou fazer predições com base nesses dados.

Neto (2004) descreve a estatística como a ciência que realiza o tratamento correto de dados que envolvem incerteza, ou uma parte da matemática aplicada, a qual fornece métodos para a coleta, a organização, a descrição, a análise e a interpretação de dados quantitativos e a utilização destes dados para a tomada de decisões.

A Estatística Descritiva é a parte da Estatística onde são descritos os dados relativos a um determinado experimento ou situação, sem que haja uma

preocupação com a análise dos dados. A Estatística Descritiva se preocupa com a descrição e organização dos dados experimentais.

A parte da Estatística que se preocupa com as técnicas de análise e interpretação dos dados é a Estatística Indutiva ou Inferencial. A partir de uma parcela dos dados, amostra, organizada e descrita pela estatística descritiva, tira-se conclusões a respeito da população (conjunto maior dos dados) por meio da Estatística Indutiva.

Alguns conceitos básicos de estatística, assim como a apresentação de cálculos estatísticos são essenciais na descrição de uma variável, os quais são apresentados na sequência.

a) Tipos de Variáveis

A variável, em Estatística, é a característica de interesse, a qual se pretende analisar. A escolha da variável (ou variáveis) depende do foco de estudo e análise estatística em questão.

Na Estatística as variáveis podem ser classificadas em *quantitativa* ou *qualitativa*. As variáveis qualitativas são aquelas que resultam em uma classificação por tipos ou atributos. Como exemplo de variáveis qualitativas pode-se citar a variável gênero, classificada em ‘masculino’ ou ‘feminino’.

As variáveis quantitativas são aquelas cujos valores são expressos em números. Esta variável pode ser subdividida em quantitativa discreta e quantitativa contínua. Variável quantitativa discreta é aquela que pode assumir apenas valores pertencentes a um conjunto enumerável. Como exemplo, o número de filhos de casais residentes em uma cidade.

Variável quantitativa contínua é aquela que pode assumir qualquer valor de um intervalo de variação. Como exemplo, a idade de pessoas residentes em uma cidade.

b) Coeficiente de Variação

O efeito da variação ou dispersão em relação à média é medido pela dispersão relativa (equação 2.1).

$$DR = \frac{Dispersão\ absoluta}{Média} \qquad (2.1)$$

Se a dispersão absoluta for o desvio padrão, a dispersão relativa é denominada coeficiente de variação (CV), que é representada pela equação 2.2.

$$CV = \frac{S}{\bar{X}} \quad (2.2)$$

Sendo: S = Desvio padrão; e $\bar{X}$ = Média.

c) Variância

A variância é a medida de variação que mais é utilizada e mede quanto os valores de uma distribuição distam de sua média, é representada pela equação 2.3.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad (2.3)$$

Sendo: X_i = Valor da variável; e $\bar{X}$ = Média.

d) Covariância

A Covariância entre duas variáveis é o valor esperado dos produtos dos desvios de cada variável aleatória em relação a sua média. Segundo Guimarães (2004) quando utilizadas variáveis bidimensionais, dizemos que a covariância é a medida de associação entre elas. O cálculo da covariância pode ainda ser pensado para a análise espacial. A covariância é dada por:

$$cov(x, y) = E\{[X - \mu x\,].[Y - \mu y]\} \quad (2.4)$$

Se analisarmos a Variável Z nas posições t e t+h temos:

$$cov[Z(t), Z(t+h)] = E[\big(Z(t) - \mu_{Z(t)}\big).Z(t+h) - \mu_{Z(t+h)}\,)] \quad (2.5)$$

Se a variável Z é estacionária, esta função poderá ser estimada por:

$$cov(Z(t), Z(t+h)) = \frac{\sum_{i=1}^{n(h)}[Z(t_i) - \bar{Z}][Z(t_i+h) - \bar{Z}]}{n(h-1)} \quad (2.6)$$

Nesse caso, a média de Z(t) será igual à média de Z(t+h).

Uma propriedade da covariância diz que se duas variáveis aleatórias são independentes, então a covariância entre elas é igual a zero. Neste caso, na análise da variável Z nas posições t e t+h, com h=1,2,...k, espera-se que o valor da covariância comece alto e depois tenda a zero, sendo que quanto maior for o valor da covariância maior será a relação espacial e para covariância zero teremos independência.

e) Desvio Padrão

É a raiz quadrada da variância (equações 2.7 e 2.8).

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad (2.7)$$

Ou

$$\sigma = \sqrt{\sigma^2} \quad (2.8)$$

O desvio-padrão é a medida mais usada na comparação de diferenças entre conjuntos de dados, por ter grande precisão. O desvio padrão determina a dispersão dos valores em relação à média (CORREA, 2003).

f) Escore Z

O escore z permite a comparação de um valor específico com a população, levando-se em conta o valor típico e a dispersão, representada pela equação 2.9:

$$Z = \left|\frac{x-\bar{y}}{\sigma}\right| \quad (2.9)$$

Onde: x é o valor avaliado; e $\overline{y}$ a média e σ o desvio padrão.

g) Regressão linear

Quando duas variáveis possuem certo grau de relacionamento, verificado pela correlação, podemos aplicar a análise de regressão que permite descrever, por meio de um modelo matemático, a relação entre duas variáveis, partindo de *n* observações das mesmas.

Para executarmos a regressão, as variáveis serão divididas em variável dependente e variável independente. Para o eixo x, indicamos a variável independente e para o eixo y, a dependente.

Dessa forma temos: $Y = \alpha + \beta x$ que é o coeficiente linear, que dá a altura em que a reta corta o eixo das ordenadas.

Neste trabalho uma das técnicas utilizadas é a de regressão linear, a qual estima os dados de temperatura das estações.

h) Correlação

A correlação de dados é descrita como a relação existente entre duas variáveis. Essa fornece um número que resume o grau de relacionamento entre elas (BISQUERRA, 2007).

Em estudos que envolvem duas ou mais variáveis é comum o interesse em conhecer o relacionamento entre elas, além das estatísticas descritivas normalmente calculadas. A medida que mostra o grau de relacionamento entre duas variáveis é chamada de coeficiente de correlação. É também conhecida como medida de associação, de interdependência, de intercorrelação ou de relação entre as variáveis (LIRA, 2004).

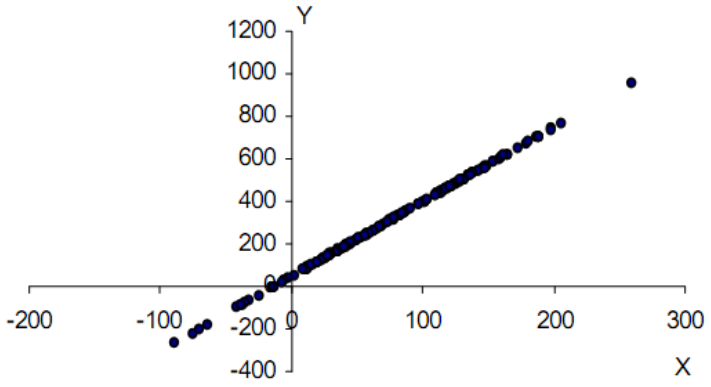
De forma geral pode-se conceituar correlação como sendo a relação ou o grau de dependência entre as duas variáveis de uma distribuição bidimensional. Quando duas variáveis estão ligadas por uma relação estatística, diz-se que existe correlação entre elas. A correlação, então, é a verificação da existência e do grau de relação entre duas (ou mais) variáveis.

Entre as variáveis podem existir diferentes formas de correlação. A mais simples e conhecida é a chamada de correlação simples, a qual envolve apenas duas variáveis. A relação entre estas duas variáveis será linear quando o valor de uma delas pode ser obtido por meio da equação da reta ($Y = \alpha + \beta x$). Nesse caso a correlação é chamada linear simples.

Caso não seja possível o ajuste das variáveis na equação da correlação linear, não significa que não exista correlação entre as variáveis, mas sim uma correlação não-linear entre elas.

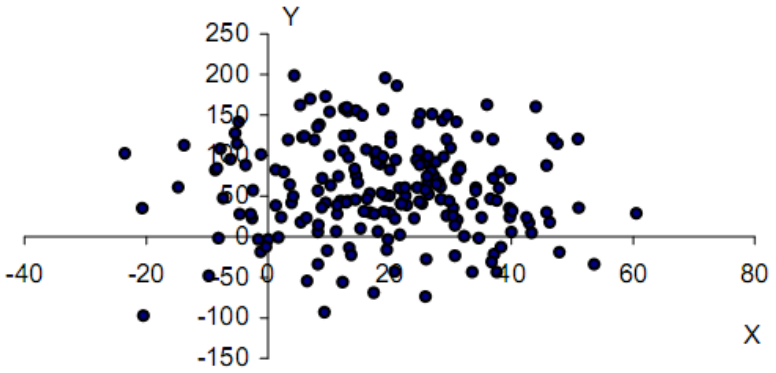
Uma das formas mais simples de verificar qual o tipo de correlação entre as variáveis utiliza-se o diagrama de dispersão, onde os dados são representados aos pares. Por meio deste diagrama pode-se identificar o tipo de correlação. Nas Figuras 2.1, 2.2 e 2.3 são apresentados três tipos de diagramas de dispersão: linear, linear nula e o não-linear, respectivamente.

Figura 2.1 - Correlação Linear perfeita entre as variáveis X e Y



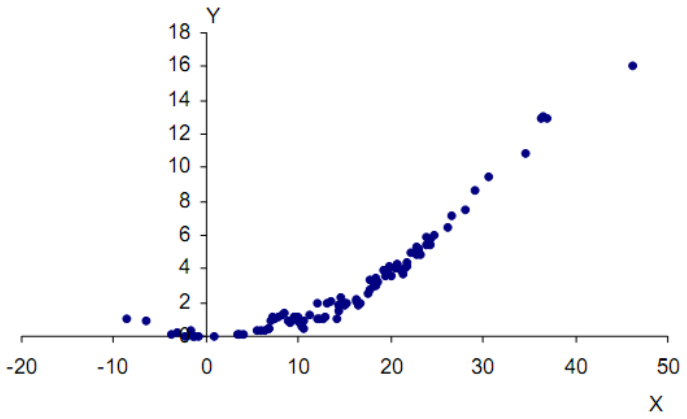
Fonte: (LIRA, 2004)

Figura 2.2 - Correlação Linear nula entre as variáveis X e Y



Fonte: (LIRA, 2004)

Figura 2.3 - Correlação Não-Linear entre as variáveis X e Y



Fonte: (LIRA, 2004)

Segundo Lira (2004), existem várias técnicas para avaliar a correlação entre duas variáveis, dentre elas destaca-se o coeficiente de correlação linear de Pearson.

O coeficiente de correlação linear de Pearson, também, é conhecido como Coeficiente de Correlação Momento Produto, e é utilizado para medir a correlação entre duas variáveis. Este coeficiente é calculado através da equação 2.10:

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (2.10)$$

O coeficiente de correlação Pearson (r ou $\hat{\rho}$) varia de -1 a 1. O sinal indica direção positiva ou negativa do relacionamento e o valor sugere a força da relação entre as variáveis. Uma correlação perfeita (-1 ou 1) indica que o escore de uma variável pode ser determinado exatamente ao se saber o escore da outra. No outro oposto, uma correlação de valor zero indica que não há relação linear entre as variáveis (FIGUEREDO et al., 2009).

Entretanto dificilmente consegue-se obter valores extremos como 0 ou 1 em experimentos práticos. Segundo Ferreira et al. (2009) a classificação e interpretação destes índices de correlação podem variar de acordo com o contexto ou pesquisador, são citadas duas classificações apresentadas nas tabelas 2.1 e 2.2.

Tabela 2.1 – Primeira Classificação de Correlação

Grau de correlação	Classificação
0,10 - 0,29	pequenos
0,30 - 0,49	médios
0,50 - 1,00	grandes

Fonte: FERREIRA et al. (2009)

Tabela 2.2 - Segunda Classificação de Correlação

Grau de correlação	Classificação
0,10 - 0,30	fraco
0,40 - 0,60	moderado
0,70 - 1,00	forte

Fonte: FERREIRA et al. (2009)

Já segundo Callegar-Jacques (2003), o coeficiente de correlação pode ser avaliado qualitativamente conforme apresentado na tabela 2.3.

Tabela 2.3 - Classificação de Correlação Callegar-Jacques

Grau de correlação	Classificação
0,00 - 0,30	fraca
0,30 - 0,60	moderada
0,60 – 0,90	forte
0,90 – 1,00	muito forte

Fonte: CALLEGAR-JACQUES (2003)

Independente da classificação adotada sabe-se que quanto mais próximo de 1 (independente do sinal) maior é o grau de dependência estatística linear entre as variáveis. No outro oposto, quanto mais próximo de zero, menor é a força dessa relação.

O índice de correlação de Pearson é utilizado, neste trabalho, a fim de eleger as estações que possuem maior relação com o ponto a ser estimado.

i) Erro Padrão

O erro padrão é uma medida da precisão da média amostral calculada. O erro padrão obtém-se dividindo o desvio padrão pela raiz quadrada do tamanho da amostra (equação 2.11).

$$S_x = \frac{\sigma}{\sqrt{n}} \tag{2.11}$$

Quando não se conhece o desvio padrão da população, usa-se o desvio padrão da(s) amostra(s), através da equação 2.12.

$$S_x = \frac{s}{\sqrt{n}} \tag{2.12}$$

Sendo: $S_x \rightarrow$ é o erro padrão;$s \rightarrow$ é o desvio padrão; e $n \rightarrow$ é o tamanho da amostra.

De forma geral, o erro padrão não é mais do que o desvio padrão da distribuição das médias das amostras, de uma população.

j) Erro Médio

O valor do Erro Médio é determinado pela equação 2.13:

$$\delta = \frac{\sum_{i=1}^{m}(x_{est(i)} - x_{obs(i)})}{m} \quad (2.13)$$

Onde δ é o Erro Médio $x_{est(i)}$ é o valor estimado na posição i, $x_{obs(i)}$ é o valor observado na posição i, m é o número de valores da amostra simulados e i representa o índice que determina a posição.

2.2 GEOESTATÍSTICA

A Geoestatística estuda as variáveis regionalizadas, ou seja, variáveis com condicionamento espacial (LANDIM, 2006). A Geoestatística baseia-se nos seguintes pressupostos:

- Ergodicidade: a esperança referente à média de todas as possíveis realizações, da variável, é igual à média de uma única realização, dentro de um domínio;
- Estacionariedade: na região em que se pretende fazer estimativas, o fenômeno é descrito como homogêneo dentro desse espaço;
- Hipótese intrínseca: as diferenças entre valores apresentam fraco incremento, isto é, as diferenças são localmente estacionárias.

Para a elaboração de um variograma, supõe-se que a variável regionalizada tenha um comportamento fracamente estacionário, onde os valores esperados, assim como sua covariância espacial, sejam os mesmos por uma determinada área. Desse modo, assume-se que os valores dentro da área de interesse não apresentem tendência que possa afetar os resultados.

Segundo Jakob (2002), a Geoestatística é a área que inclui uma variedade de técnicas de estimação, como inverso do quadrado da distância, análise do vizinho mais próximo, e krigagem linear e não-linear. É mais comumente usada para identificar e mapear padrões espaciais da superfície terrestre. Pode ser usada para determinar se existe autocorrelação espacial entre dados de pontos. Para isso, a função mais comum utilizada é o (semi) variograma.

Segundo Druck (2002), a geoestatística envolve um conjunto de procedimentos de análise e inferência dos fenômenos espaciais, que apresentem uma dependência espacial expressa. Por exemplo, numa função de autocorrelação no espaço. Esses procedimentos utilizam um modelo de autocorrelação espacial definido “a priori” e objetivam representar a variabilidade espacial de um atributo, considerado em uma superfície contínua. São, portanto, procedimentos que podem ser aplicados a vários tipos de fenômenos, tais como concentração de poluentes e variação do teor de zinco no solo, apesar de ter sua origem na área de mineração.

Mediante estas técnicas, dentre as quais se destacam a krigagem e a simulação estocástica, é possível calcular um valor de uma dada propriedade, para cada centro da célula de uma malha tridimensional. Esse valor está condicionado aos dados existentes e a uma função de correlação espacial entre esses dados.

Em várias áreas das Ciências da Terra, as variáveis não apresentam um padrão de distribuição requerido pela estatística clássica, como normalidade e independência dos dados. Os modelos da estatística clássica estão, geralmente, voltados para a verificação da distribuição de frequência dos dados, enquanto a geoestatística incorpora a interpretação da distribuição estatística, assim como a correlação espacial das amostras. Este aspecto da geoestatística está associado com a distribuição estatística dos dados no espaço.

A Geoestatística define um conjunto de procedimentos matemáticos que permite que se reconheça e descreva relacionamentos espaciais existentes. Neste processo, admite-se que a posição de uma amostra é tão importante quanto o valor medido, ou seja, assume-se o princípio de que tudo está relacionado a tudo, mas o que se encontra mais próximo está mais relacionado.

Assim, os métodos geoestatísticos fornecem um conjunto de técnicas para entender a uma aparente aleatoriedade dos dados, mas com possível estruturação espacial, estabelecendo, desse modo, uma função de correlação espacial. Esta função representa a base da estimativa da variabilidade espacial.

2.2.1 Variograma

O variograma é a descrição matemática do relacionamento entre a variância de pares de observações (pontos) e a distância, separando estas observações (h). A

autocorrelação espacial pode, então, ser usada para fazer melhores estimativas para pontos não amostrados (JAKOB, 2002).

O variograma é considerado ferramenta básica, que permite descrever quantitativamente a variação no espaço de um fenômeno regionalizado.

A função variograma 2γ(h) é definida como sendo a esperança matemática do quadrado da diferença (equação 2.14), entre os valores de pontos no espaço, separados por uma distância h.

$$2\gamma(h) = E\{[Z(x+h) - Z(x)]^2\}$$

$$2\gamma(h) = \frac{1}{n}\sum_{i=1}^{n}[Z(x+h) - Z(x)]^2 \qquad (2.14)$$

Sendo: n – números de pares de pontos separados por distância h; Z(x) - valor da variável regionalizada no ponto x; e Z(x + h) valor da variável regionalizada no ponto x +h.

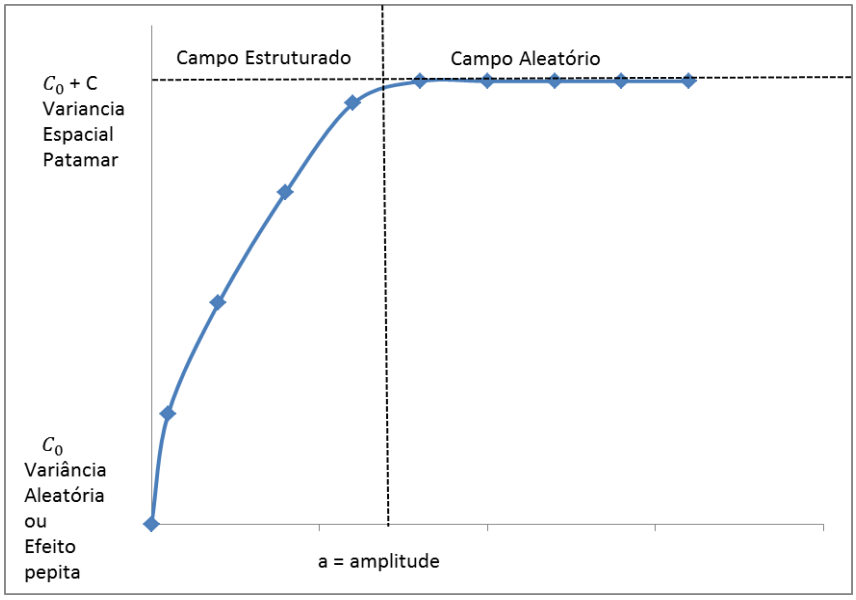
A interpretação do variograma permite obter parâmetros que descrevem o comportamento espacial das variáveis regionalizadas.

Uma feição resultante da análise dos parâmetros do variograma experimental é a zona de influência: qualquer valor de Z(x) estará correlacionado com outros valores Z(x+h) que estiverem dentro de um raio "a" de x. Esta correlação, ou a influência de um valor em outro, decresce conforme Z(x+h) aproxima-se de "a".

O variograma é utilizado para calcular os valores de semivariância, para uma dada distância, os quais são necessários para a organização do sistema de equações de krigagem.

O variograma substitui a distância euclidiana “h” pela distância “2γ(h)”, atributo específico do local em estudo. A distância dada pelo variograma mede o grau médio de similaridade entre um valor não amostrado e um valor conhecido vizinho. Os componentes do variograma são apresentados na Figura 2.4.

Figura 2.4 - Componentes do variograma



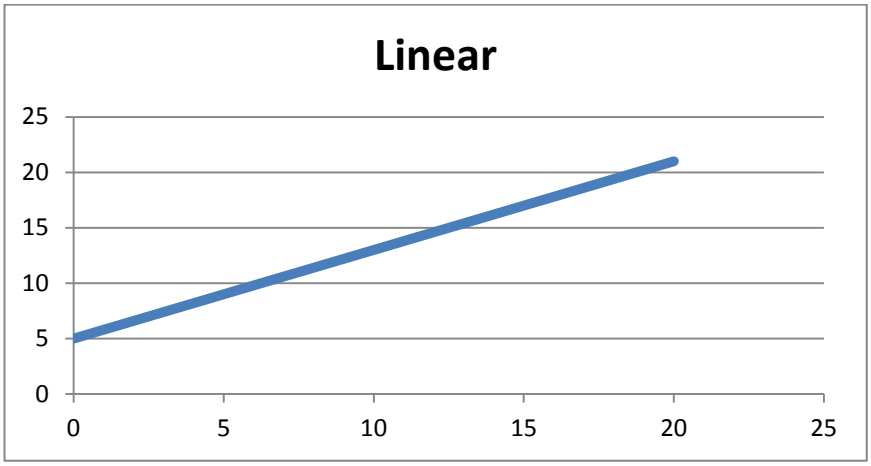
Fonte: O autor

A amplitude é distância a partir da qual as amostras passam a ser independentes. Ela reflete o grau de homogeneidade entre as amostras e quanto maior a amplitude maior a homogeneidade entre as amostras. Dessa forma, o variograma dá um significado preciso da noção tradicional de zona de influência. A amplitude (a) separa o campo estruturado, amostras correlacionáveis, do campo aleatório, amostras independentes.

O patamar é o valor da variância no qual o variograma se estabiliza, no campo aleatório. O Efeito pepita ou variância aleatória é o valor da função variograma na origem (h = 0). O C_0 deveria ser 0, pois duas amostras tomadas no mesmo ponto (h=0) deveriam ter os mesmos valores.

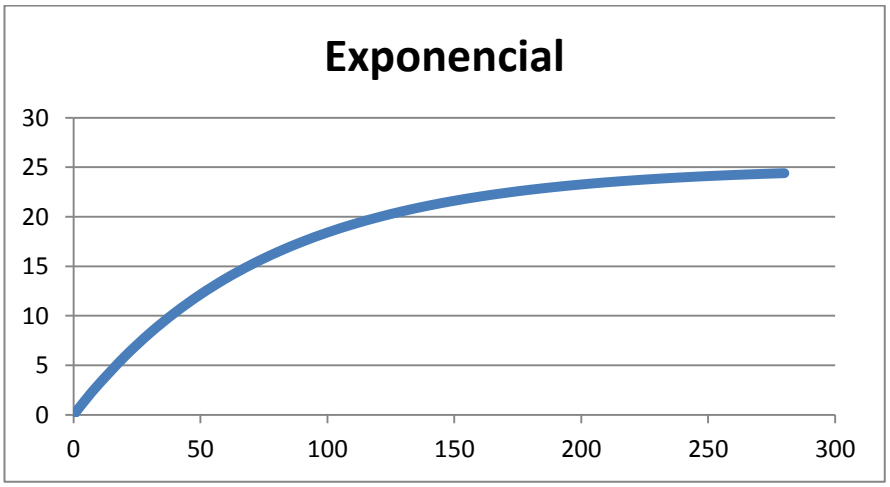
Variância Espacial é dada pela diferença entre a variância e o efeito pepita. De acordo com Jakob (2002), quanto aos modelos de variograma, os mais comuns são: Linear, Exponencial, Esférico e Gaussiano, os quais são representados nas Figuras 2.5, 2.6, 2.7 e 2.8, respectivamente.

Figura 2.5 - Modelo de variograma linear



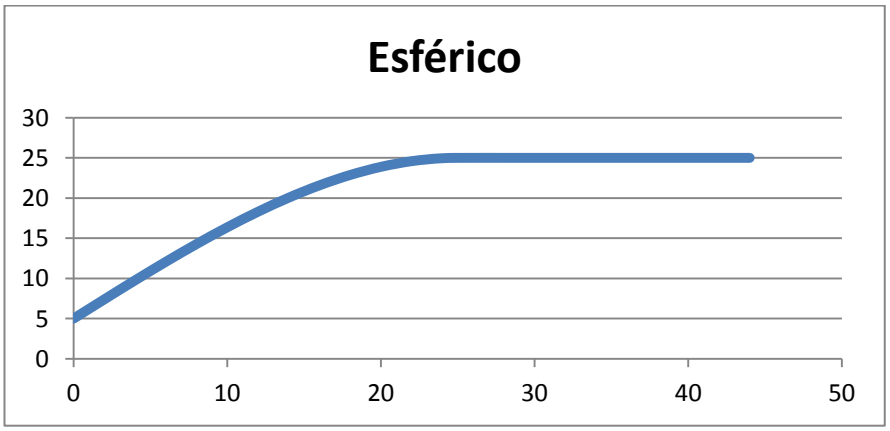
Fonte: O autor

Figura 2.6 - Modelo de variograma exponencial



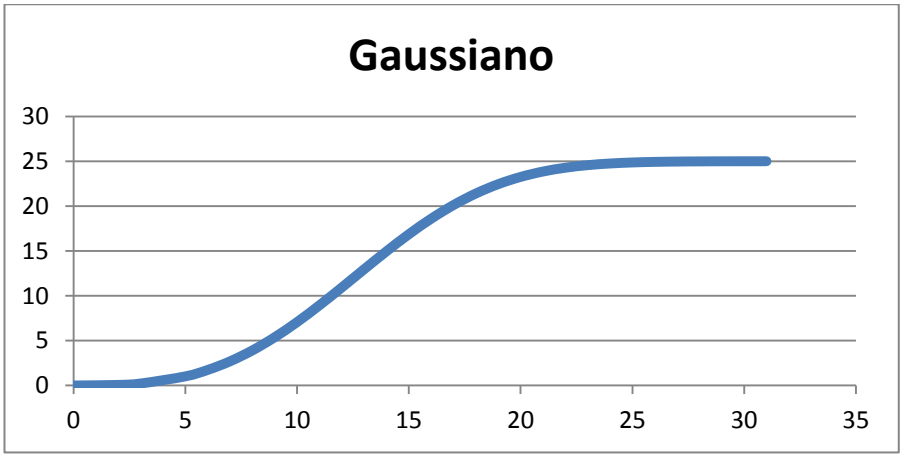
Fonte: O autor

Figura 2.7 - Modelo de variograma esférico



Fonte: O autor

Figura 2.8 - Modelo de variograma gaussiano



Fonte: O autor

Ainda segundo Jakob (2002) a parte importante do variograma é sua forma próxima à origem, uma vez que são os pontos mais próximos que possuem maior peso no processo de interpolação.

2.2.2 Semivariograma

Segundo Guimarães (2004) o semivariograma é definido pela equação 2.15:

$$\gamma(h) = \frac{1}{2}\{Var[Z(t) - Z(t+h)]\} \quad (2.15)$$

Sendo: Var[Z(t) –Z(t+h)] a variância dos dados separados por uma distância h. A variância é dividida por dois, então utiliza-se o prefixo "semi" para distinguir da variância, por conta disto, vem o nome semivariância para γ(h) e semivariograma para o gráfico de γ(h), em função de h.

Sob a suposição de tendência zero, temos: E[Z(t+h)] = E[Z(t)] e, portanto:

$$\gamma(h) = \frac{1}{2}\{E[Z(t+h) - Z(t)]^2\} \quad (2.16)$$

e uma estimativa de γ(h) chamada de $\hat{\gamma}$(h) é dada por:

$$\hat{\gamma}(h) = \frac{\sum_{i=1}^{n(h)}[Z(t+h)-Z(t)]^2}{2n(h)} \quad (2.17)$$

Sendo: n(h) = número de pares separados pela distância h.

Analisando a expressão da função semivariância, pode-se concluir que

quanto mais próximos estiverem os pontos amostrados, maior será a semelhança entre eles e, portanto, menor a semivariância. De modo similar, conclui-se que quanto mais distantes estiverem os pontos amostrados, menores serão as semelhanças e por consequência maior será dispersão (variância).

Teoricamente, temos para a distância h=0 a semivariância γ(0) = 0 com a semivariância γ(h) aumentando com o incremento de h, até que atinja um valor constante para γ(h) que corresponde às variações aleatórias.

A seguir são descritos os parâmetros do semivariograma, segundo as definições de Druck (2002), os quais podem ser observados na Figura 2.9.

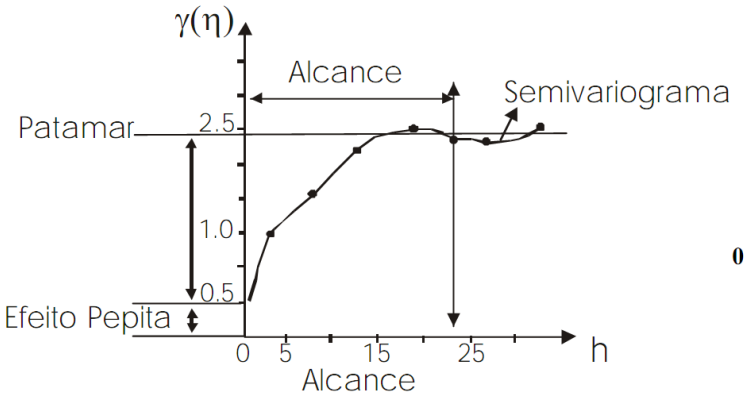
O alcance ou alcance da dependência espacial (a) é a distância, dentro da qual as amostras apresentam-se correlacionadas espacialmente.

O patamar (C) é o valor do semivariograma correspondente a seu alcance (a). Deste ponto em diante, considera-se que não existe mais dependência espacial entre as amostras.

O efeito pepita (C_0), por definição, γ(0)=0. Entretanto, na prática, à medida que h tende para zero, γ(h) se aproxima de um valor positivo chamado Efeito Pepita (C_0). O valor de C_0 revela a descontinuidade do semivariograma para distâncias menores do que a menor distância entre as amostras. Parte dessa descontinuidade pode ser, também, devida a erros de medição. No entanto, é impossível quantificar se a maior contribuição provém dos erros de medição ou da variabilidade de pequena escala não captada pela amostragem.

A Contribuição (C_1) é a diferença entre o patamar (C) e o efeito pepita (C_0).

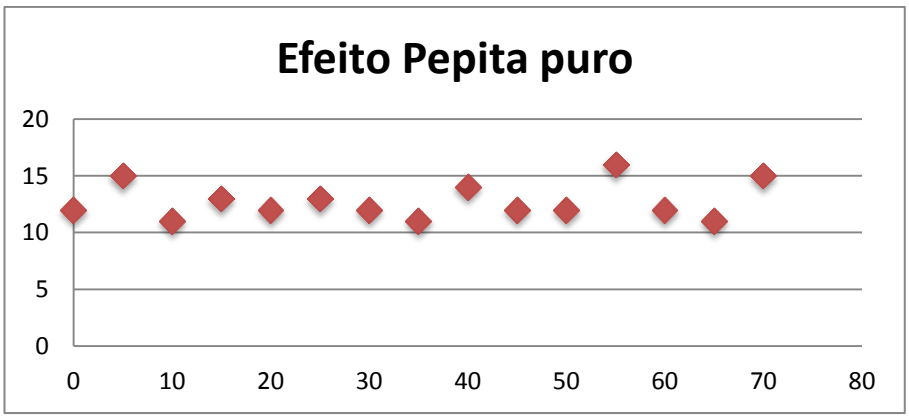
Figura 2.9 - Parâmetros do semivariograma



Fonte: DRUCK (2002)

No caso do semivariograma for constante e igual ao patamar para qualquer valor de h, caracteriza o efeito pepita puro (Figura 2.10), representando assim a ausência total de dependência espacial. Nesse caso, a dependência espacial, caso exista, será manifestada na distância ou no tempo menor do que o menor espaçamento entre amostras.

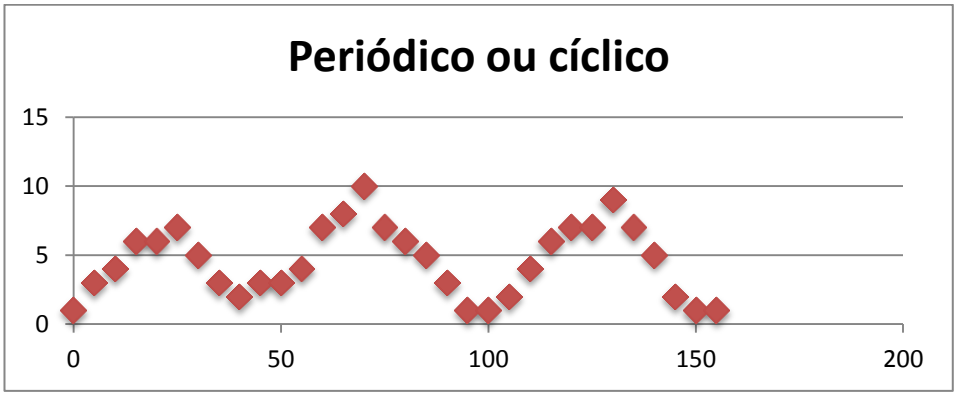
Figura 2.10 - Efeito pepita pura



Fonte: O autor

Outro tipo de semivariograma é aquele cuja semivariância apresenta flutuações, chamado de periódico ou cíclico (Figura 2.11) e indica uma periodicidade nos dados explicada por algum fator conhecido e analisada por meio da densidade espectral.

Figura 2.11 - Semivariograma periódico ou cíclico

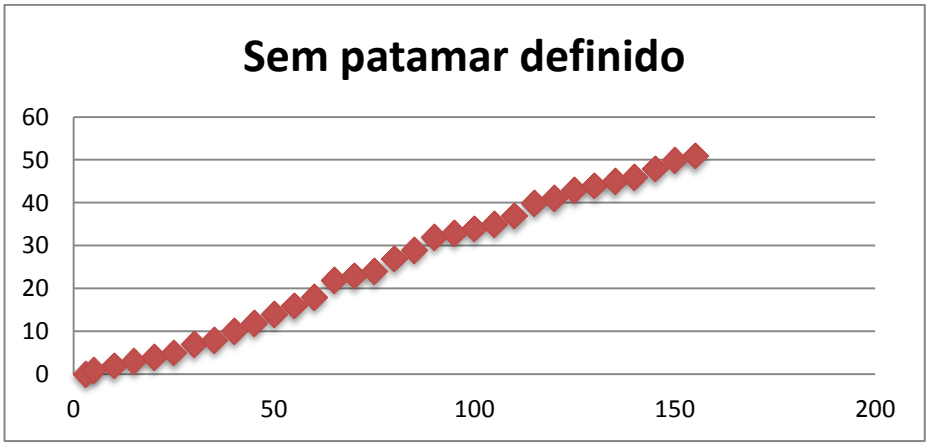


Fonte: O autor

Existe, ainda, o tipo de semivariograma, onde as semivariâncias crescem, sem limite, para todos os valores de h, ou seja, semivariogramas sem patamar

definido (Figura 2.12). Este semivariograma indica que a hipótese de estacionaridade de segunda ordem não foi atendida e, provavelmente, estamos trabalhando com a hipótese intrínseca, fenômeno com capacidade infinita de dispersão, indicando que a máxima distância h, entre as amostras, não foi capaz de exibir toda a variância dos dados e, provavelmente, existe tendência dos dados para determinada direção.

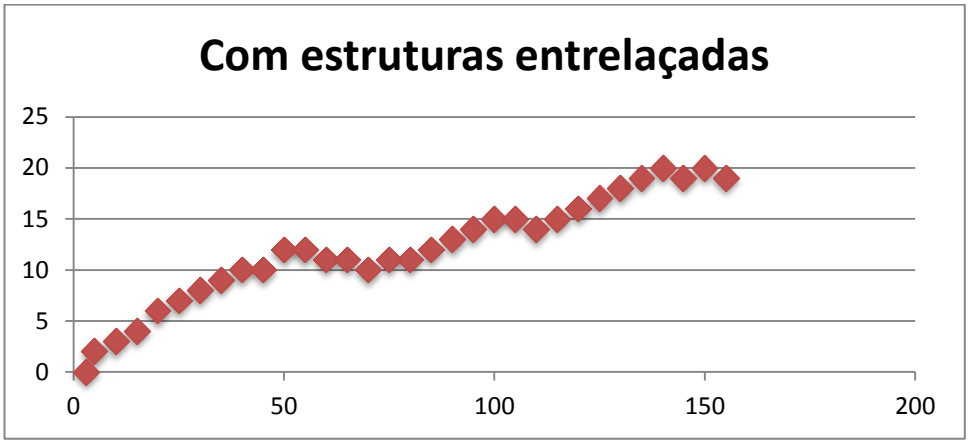
Figura 2.12 - Semivariograma sem patamar definido



Fonte: O autor

O semivariograma, com mais de uma estrutura de variância, é chamado de semivariograma com estruturas entrelaçadas ou semivariograma imbricado (Figura 2.13). Nesse caso, uma explicação prática pode estar associada ao fato de uso de mais de uma população, ou seja, até uma determinada distância estamos utilizando uma determinada população e a partir desta distância outra população.

Figura 2.13 - Semivariograma com estruturas entrelaçadas



Fonte: O autor

De acordo com Guimarães (2004), com relação à dependência espacial da variável, a qual está sendo estudada, pode-se classificá-la como segue:

- Variável com forte dependência espacial – se o efeito pepita for menor ou igual a 25% do patamar $\left(\frac{C_0}{C_0+C} < 0{,}25\right)$
- Variável com moderada dependência espacial – se o efeito pepita representar entre 25% e 75% do patamar $\left(0{,}25 \leq \frac{C_0}{C_0+C} \leq 0{,}75\right)$
- Variável com fraca dependência espacial – se a relação entre efeito pepita e patamar $\left(0{,}75 < \frac{C_0}{C_0+C} < 1{,}00\right)$
- Variável independente espacialmente – se a relação entre efeito pepita e patamar for igual a 100%, neste caso temos o semivariograma, com efeito pepita puro $\left(\frac{C_0}{C_0+C} = 1{,}00\right)$.

Quanto aos conceitos de Isotropia e Anisotropia pode-se dizer, de acordo com Guimarães (2004), sendo *h* um vetor e, consequentemente, o semivariograma depende da magnitude e da direção de *h*. Quando o semivariograma é idêntico para qualquer direção de *h* ele é chamado de isotrópico e quando o semivariograma apresenta os parâmetros *C*, C_0, *a* e/ou modelo diferenciado dependendo da direção de *h*, ele é chamado anisotrópico. Neste trabalho será utilizado o conceito de isotropia.

A escolha do modelo de semivariograma é um dos aspectos importantes da geoestatística. Todos os cálculos da geoestatística dependem do modelo de semivariograma ajustado e, consequentemente, se o modelo ajustado não for

apropriado, todos os cálculos seguintes conterão erros que poderão afetar as inferências. Portanto, o ajuste de semivariograma é uma fase crucial na análise geoestatística e deve receber uma atenção especial (GUIMARÃES, 2004).

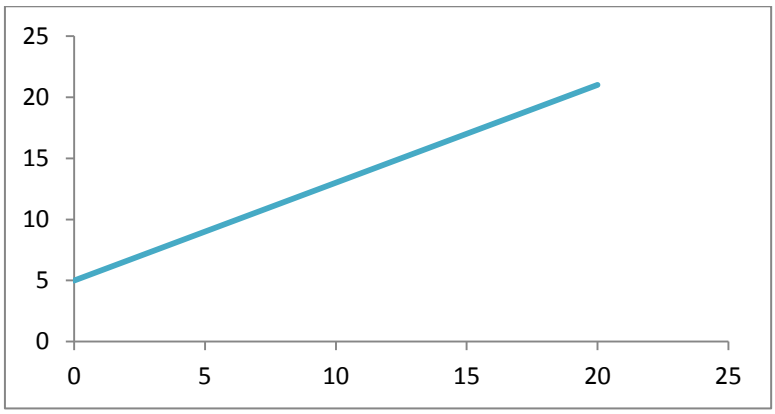
Definindo C_0 como efeito pepita, C_0 + C como patamar e a como alcance, os principais modelos de semivariogramas utilizados na geoestatística são:

- Modelo linear com patamar

$$\gamma(h) = \{ C_0 + \frac{C}{a} h\} \quad se \quad 0 \leq h \leq a$$

$$\gamma(h) = \{ C_0 + C\} \quad se \ h > a \qquad (2.18)$$

Figura 2.14 - Semivariograma Modelo Linear



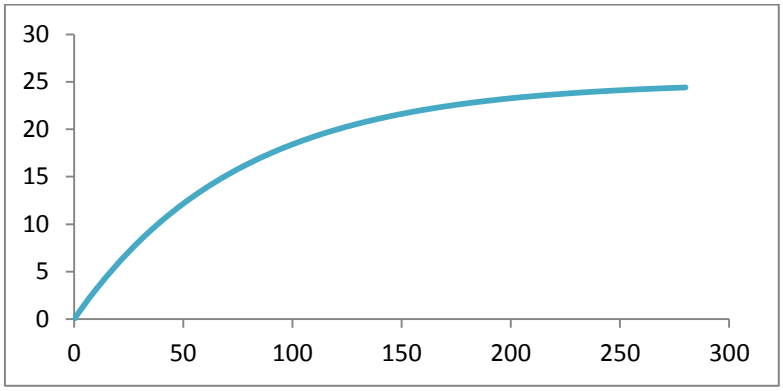
Fonte: O autor

- Modelo esférico

$$\gamma(h) = \{ C_0 + C \left[\frac{3}{2} \left(\frac{h}{a}\right) - \frac{1}{2} \left(\frac{h}{a}\right)^3 \right] \quad se \ 0 \leq h \leq a$$

$$\gamma(h) = \{ C_0 + C\} \quad se \ h > a \qquad (2.19)$$

Figura 2.15 - Semivariograma Modelo Esférico



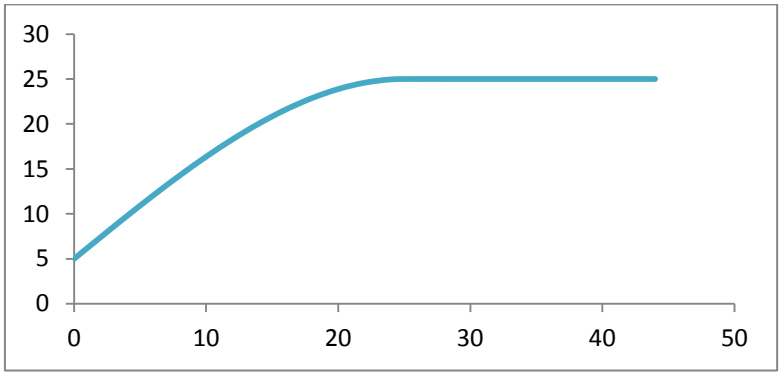
Fonte: O autor

- Modelo exponencial:

$$\gamma(h) = C_0 + C\left[1 - e^{[-3(h/a)]}\right] \quad se \;\; 0 < h < d \qquad (2.20)$$

Neste modelo e no modelo gaussiano, *d* é a distância máxima na qual o semivariograma é definido e nesses modelos o patamar *a* é atingido, apenas, assintoticamente O parâmetro a é determinado visualmente como a distância após a qual o semivariograma se estabiliza.

Figura 2.16 - Semivariograma Modelo Exponencial

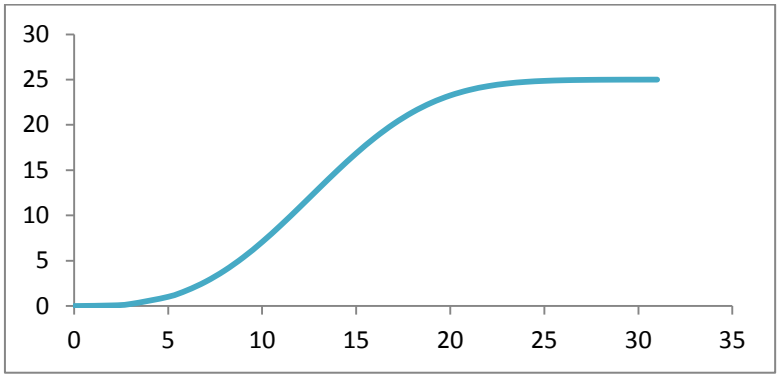


Fonte: O autor

- Modelo gaussiano

$$\gamma(h) = C_0 + C\left[1 - e^{[-3(h/a)^2]}\right] \quad se \;\; 0 < h < d \qquad (2.21)$$

Figura 2.17 - Semivariograma Modelo Gaussiano



Fonte: O autor

2.2.3 Interpolação

Interpolação é o método que permite construir um novo conjunto de dados a partir de um conjunto discreto de dados pontuais previamente conhecidos. Através da interpolação, pode-se construir uma função que aproximadamente se "encaixe" nestes dados pontuais, conferindo-lhes, então, a continuidade desejada. Segundo Burrough e Mcdonnell (1998) a interpolação é necessária quando:

- Uma superfície contínua é representada por modelos de dados que diferem do requerido. Por exemplo, as transformações de superfícies contínuas de um tipo para outro, dados irregulares para grade regular;
- Computação de elevação (Z) de dados pontuais;
- Computação de elevação (Z) de uma grade retangular original (denominado "*gridding*");
- Computação de locais (X,Y) de pontos ao longo de contornos (em interpolação de contorno);
- Densificação ou aumento de grades retangulares (denominado reamostragem).

Logo, a interpolação espacial é o procedimento para se estimar valores de propriedades de locais não amostrados, baseando-se em valores de dados observados em locais conhecidos. Os interpoladores são distinguidos em globais ou locais; exatos ou suavizantes; e determinísticos ou estocásticos (BURROUGH, 1986).

a) *Interpoladores Globais*

Os Interpoladores globais consideram todos os pontos da área amostrada, permitindo interpolar o valor da função em qualquer ponto dentro do domínio dos dados originais, já que determinam apenas uma função que é mapeada através de toda a região. A adição ou remoção de um valor tem consequências no domínio de definição da função. Existe ainda uma tendência dos algoritmos de interpolação global a gerar superfícies mais suaves, com mudanças menos bruscas.

b) *Interpoladores Locais*

Os interpoladores locais são funções definidas para porções determinadas, portanto a alteração de um valor afeta localmente os pontos próximos ao mesmo.

c) *Interpoladores Exatos*

Interpoladores exatos, de forma geral, são utilizados quando se tem certeza dos valores dos pontos, nos quais a interpolação está baseada. Eles sempre honram os dados, de maneira que após o processo de interpolação não há presença de resíduos, ou seja, a predição sobre os locais amostrados vai ser igual ao próprio valor amostrado.

d) *Interpoladores Suavizantes*

Os interpoladores suavizantes ou *“smoothing”*, ao contrário dos interpoladores exatos, são utilizados quando há incerteza sobre os valores dos pontos amostrados, geralmente provenientes de locais que sofrem variações ou flutuações rápidas. Esses interpoladores produzem suavização das curvas da superfície gerada, fazendo com que possíveis erros presentes nos dados tendam a ser minimizados.

e) *Interpoladores Estocásticos*

Os interpoladores estocásticos fazem uso da teoria da probabilidade, e incorporam critérios estatísticos na determinação do peso atribuído aos pontos amostrais, para o cálculo das interpolações.

f) Interpoladores Determinísticos

Os interpoladores determinísticos não fazem uso da probabilidade. Para calcular a medida de uma grandeza no espaço, eles geram uma combinação linear dos valores amostrados, baseando-se apenas na geometria da distribuição espacial dos dados amostrados.

Segundo Miranda (2005), o processo de interpolação é constituído de duas partes, sendo a primeira, a definição de um relacionamento de vizinhança, e a segunda, a definição de qual método calculará os valores desconhecidos.

Os métodos de interpolação são utilizados para estimar dados a partir de um conjunto de amostras, previamente, coletadas. Ou seja, o conceito básico é gerar novos conjuntos de dados a partir de um conjunto discreto de dados conhecidos. Por meio desse método pode-se elaborar uma função que aproxima os dados pontuais obtidos, a partir de uma amostragem ou experimento. Por conta disso, este método é utilizado por várias áreas da engenharia e da ciência em geral.

Dentre os métodos de interpolação destacam-se: Inverso da distância, Vizinho mais próximo, *Spline* e Geoestatística/kriging.

a) Inverso da distância

Também chamado de IDW (*Inverse Distance Weighted*), o inverso ponderado da distância é um método puramente matemático. Segundo Miranda (2005), este método estima um valor para um local não amostrado como uma média dos valores dos dados dentro de uma vizinhança. O cálculo da média é ponderado pela distância entre o ponto a ser interpolado e seus vizinhos. Destaca-se que o peso da distância é ajustado por um expoente, isso implica que, quanto maior expoente, maior será a influência da distância.

A ponderação do inverso das distâncias implementa, explicitamente, o pressuposto de que as coisas mais próximas entre si são mais parecidas, do que as mais distantes. Para predizer um valor para algum local não medido, o IDW usará os valores amostrados à sua volta, que terão um maior peso do que os valores mais distantes, ou seja, cada ponto possui uma influência no novo ponto, que diminui na medida em que a distância aumenta (JAKOB & YOUNG, 2006).

b) *Mínima curvatura ou Spline*

Segundo Andriotti (2009), essa é uma técnica de interpolação que utiliza um polinômio, para gerar uma superfície que minimize a curvatura da mesma, resultando em uma superfície suavizada, passando através dos pontos amostrados, reproduzindo os valores da variável. Isso não ocorre sempre, não podendo, desta forma, ser considerado um interpolador exato.

c) *Krigagem*

A krigagem é um método de interpolação que utiliza de geoestatística, possuindo, em sua base conceitual, dois fundamentos importantes, variáveis regionalizadas e funções aleatórias (MIRANDA, 2005). Na krigagem, o processo assemelha-se ao da interpolação por média ponderada do inverso da distância, onde nesse método, os pesos são determinados a partir de uma análise espacial, baseados no semivariograma.

A krigagem produz a melhor estimativa linear, não viciada dos dados, de um atributo em um local não amostrado, com a modelagem do variograma. “A krigagem ordinária é linear porque suas estimativas são combinações lineares ponderadas dos dados disponíveis; é não-viciada porque busca o valor de erro ou resíduo médio igual a 0; e é melhor porque minimiza a variância dos erros.” (Isaaks e Srivastava, 1989)

Existem diversos tipos de krigagem, com suas especificidades, como a simples, a ordinária, a universal, a pontual, a de blocos e a co-krigagem.

d) *Vizinho mais próximo*

De acordo com Franke (1982), o algoritmo de vizinho mais próximo é o método mais simples, e tem como principal característica, assegurar que o valor interpolado seja um dos valores original, ou seja, não gera novos valores. O produto final desse interpolador é caracterizado por um efeito de degrau.

2.3 DADOS ESPACIAIS E DADOS METEOROLÓGICOS

O armazenamento de dados espaciais evoluiu nos últimos anos. Cada vez mais novos recursos são adicionados aos novos SGBD (Sistema Gerenciador de Banco de Dados), uma delas é a manipulação de dados espaciais, o que permite que aplicações com recursos espaciais se tornem cada vez mais comuns. Estima-se que cerca de 80% das informações contidas nos bancos de dados contém aspectos espaciais (endereço, cidade, local de armazenamento, etc...), portanto, da importância deste recurso.

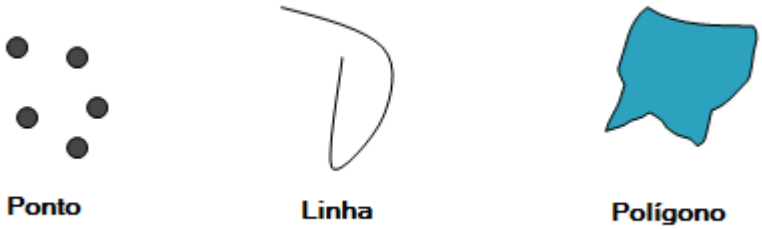
O armazenamento de dados meteorológicos em um banco de dados espacial permite que os mesmos sejam manipulados de forma rápida e eficiente, permitindo ainda a integração e cruzamento com outras informações.

a) Dados espaciais e bancos de dados espaciais

Segundo Ferreira (2003), dado espacial ou geográfico é um termo usado para representar fenômenos do mundo real através de duas componentes: (a) sua localização geográfica, ou seja, sua posição em um sistema de coordenadas conhecido; e (b) seus atributos descritivos, como por exemplo, cor, custo, pH, etc. A localização geográfica é representada por coordenadas em um sistema de coordenadas específico, onde uma coordenada é um número que representa uma posição relativa a um ponto de referência.

Um dado espacial pode ser representado por dois modelos de dados distintos: vetorial ou matricial (raster) (BURROUGH & MCDONNEL, 1998). O modelo de dados vetorial é utilizado para representar o espaço como um conjunto de entidades discretas (geo-objetos ou objetos geográficos) definidas por uma unidade (ponto, linha ou polígono, Figura 2.18) geograficamente referenciada e por seus atributos descritivos. Por exemplo, as bacias hidrográficas do Brasil podem ser representadas por objetos geográficos, onde cada bacia é representada por um polígono que define o seu limite, cada rio por uma linha e cada usina hidrelétrica por um ponto. Além da localização, cada objeto geográfico tem seus atributos descritivos, como a população de cada bacia, a potência gerada em cada usina, os nomes dos rios, etc.

Figura 2.18 - Representação dos dados espaciais Ponto, Linha e Polígono.



Fonte: O autor

A partir de 1980 os SGBD relacionais tornaram-se bastante populares por oferecerem recursos que garantiam a qualidade, segurança e facilidade de manipulação dos dados. Em seguida novos recursos foram adicionados a estes SGBD em especial a capacidade de criar e manipular objetos, de modo que estes SGBD passaram a ser denominados de bancos de dados objeto-relacional (SGBDOR). A capacidade de armazenar novos tipos de dados foi fundamental para que novas funcionalidades fossem adicionadas aos SGBD, dentre elas a capacidade de armazenar dados espaciais.

O SGBD, com extensão espacial, é uma ferramenta utilizada para armazenamento e manipulação dos dados geográficos. A diferença entre um SGBD espacial e um convencional é sua capacidade de armazenar tanto os atributos descritivos dos dados geográficos, quanto às diferentes geometrias dos mesmos.

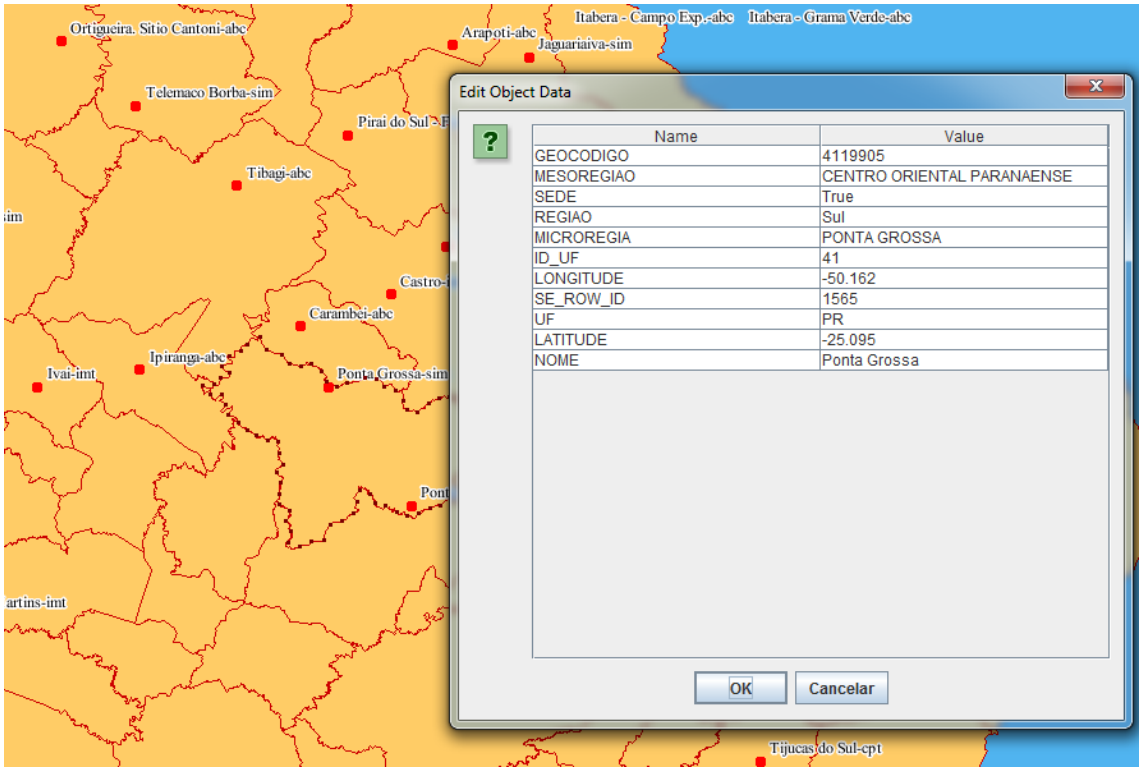
Pelo fato do SGBD armazenar a localização geográfica dos pontos a serem estudados, é possível determinar a distância entre estes pontos. Esta é a técnica utilizada neste trabalho para determinação de distância entre os pontos de interesse. Um exemplo da visualização de dados espaciais utilizados neste trabalho é apresentado na Figura 2.19.

Atualmente os recursos de manipulação de dados espaciais estão disponíveis em SGBD gratuitos como: tais como PostgreSQL, DB2 Express-C e MySQL, permitindo que as funcionalidades espaciais tornem-se cada vez mais comuns.

Dentre as vantagens em armazenar dados espaciais em um SGBD destacam-se:

- Permitir que os dados geográficos sejam armazenados, acessados e gerenciados de forma integrada com os outros dados, além de permitir a integração entre os dados espaciais e os convencionais;
- Permitir o uso da Linguagem SQL para análise de dados espaciais e da lógica de negócios;
- Permitir o uso de tipos padrões de dados espaciais para descrição de informações espaciais;
- Prover funções para consultas de dados espaciais;
- Permitir estender as aplicações atuais no SGBD, para que possua suporte a dados espaciais; e
- Permitir o uso em conjunto com vários Sistemas de Informação Geográfica.

Figura 2.19 - Exemplo da ferramenta representando os dados espaciais



Fonte: O autor

Exemplo de recuperação de coordenadas:

```
  SELECT ID_AREA_MONITORAMENTO as ID,
         NOME_AREA_MONITORAMENTO as NOME_AREA,
         VARCHAR(db2gse.ST_AsText(LOCALIZACAO), 50) AS WKTGEOM
    FROM DMET.AREA_MONITORAMENTO;
```

```
ID  NOME_AREA                          WKTGEOM
--- ---------------------------------- --------------------------------
  1 Telemaco Borba-sim                 POINT(-50.61666831 -24.33324931)
  2 Sao Miguel Arcanjo-imt             POINT(-48.1645 -23.8516)
  3 Capao Bonito - iac                 POINT(-48.33926666 -24.00399445)
  4 Itapeva-imt                        POINT(-48.8853 -23.9814)
  5 Itabera - iac                      POINT(-49.13708611 -23.86204722)
  6 Paranapanema - iac                 POINT(-48.72659167 -23.38755833)
  7 Santa Terezinha do Itaipu-cpt      POINT(-54.45536892 -25.45281372)
  8 Ivinhema-imt                       POINT(-53.8166 -22.3)
  9 Cacador-imt                        POINT(-50.8353 -26.8193)
 10 Curitibanos-imt                    POINT(-50.6042 -27.2886)
 11 Dionizio Cerqueira-imt             POINT(-53.6328 -26.2864)
 12 Indaial-imt                        POINT(-49.2683 -26.9164)
 13 Itapoa-imt                         POINT(-48.6417 -26.0813)
```

Exemplo de cálculo de distância até a coordenada (-54.26293841 -25.17503196):

```
   SELECT a.ID_AREA_MONITORAMENTO as ID,
          a.NOME_AREA_MONITORAMENTO,
          cast(DB2GSE.ST_Distance(e.LOCALIZACAO, a.LOCALIZACAO, 'KILOMETER')
               AS DECIMAL(10,4)) as distancia
     FROM dmet.AREA_MONITORAMENTO A,
          dmet.DADOS_METEOROLOGICOS D,
          TABLE(DMET.ESTACAO_ERRO ('POINT(-54.26293841 -25.17503196)')) E
    WHERE A.ID_AREA_MONITORAMENTO = D.ID_AREA_MONITORAMENTO
      AND d.data_hora = '2009-02-28 00:00:00'
      AND cast(DB2GSE.ST_Distance(e.LOCALIZACAO, a.LOCALIZACAO, 'KILOMETER')
               AS DECIMAL(10,4)) < 100
      AND cast(DB2GSE.ST_Distance(e.LOCALIZACAO, a.LOCALIZACAO, 'KILOMETER')
               AS DECIMAL(10,4)) > 0;
```

```
 ID  NOME_AREA_MONITORAMENTO     DISTANCIA
 --- --------------------------- ---------
  35 Toledo                        70.1835
  50 Santa Helena                  39.1950
  73 Cascavel                      78.8814
  81 Foz do Iguacu                 52.2477
  82 Foz do Iguacu2                34.2996
  86 Marechal Candido Rondon       97.9813
  91 Salto Caxias                  87.1249
 110 Planalto                      79.6802
```

b) Dados Meteorológicos

O acompanhamento dos dados meteorológicos tem sua importância devido às mudanças climáticas e pelo aumento de emissões de gases provenientes das atividades humanas.

A Meteorologia é o estudo do clima, o estado geral instantâneo da atmosfera em um determinado lugar e tempo. A meteorologia é descrita da medição direta de determinadas propriedades atmosféricas tais como umidade, temperatura, precipitação, descargas elétricas, entre outros (AHRENS, 2006; MENDONÇA. et al., 2007; ROHLI. et al., 2003).

Os dados meteorológicos auxiliam os agricultores na tomada de decisão, permitindo determinar a época ideal de colheita, previsão de geadas, enchentes e secas. Um período de chuva prolongado pode causar enchentes, inundando grandes extensões de terras agrícolas, assim como a falta de chuva pode causar seca, prejudicando colheitas.

As informações agrometeorológicas auxiliam no planejamento agrícola. No entanto, nem sempre essas informações encontram-se disponíveis para os usuários. Falhas nos equipamentos, erros de leitura, ou até mesmo dados perdidos durante sua transmissão, levam as estações meteorológicas a apresentarem períodos sem observações, ou com observações inconsistentes. Para solucionar esses problemas de falhas e inconsistências têm-se estudado várias técnicas para resolução desta questão.

3 MATERIAIS E MÉTODOS

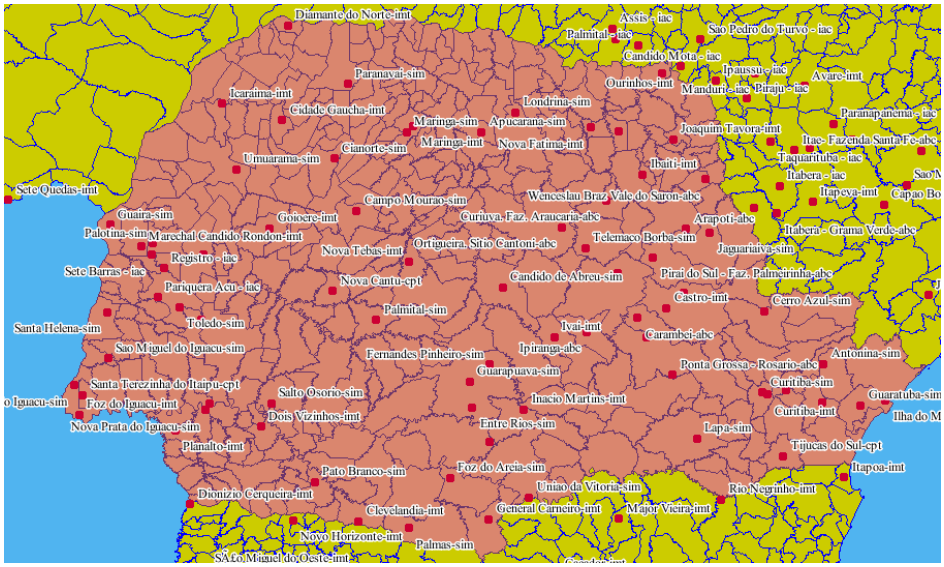
A necessidade de correção de dados meteorológicos se deve a ausência ou inconsistência dos dados. A metodologia utilizada na pesquisa para minimizar o problema é mediante o uso de técnicas estatísticas e computacionais para identificar valores inconsistentes e estimar novos valores. Para os dados faltantes, as técnicas de estimativas foram utilizadas, visto que a ausência dos dados é facilmente identificada. A metodologia proposta consiste em obter dados coletados das estações meteorológicas, inseri-las em um banco de dados com capacidade espacial, e aplicar uma série de técnicas para validação e identificação de erros, e estimar os valores, utilizando-se de técnicas estatísticas e geoestatísticas, e analisando os resultados obtidos por cada uma destas técnicas.

3.1 BASE DE DADOS METEOROLÓGICOS

Para a realização dos experimentos foi utilizado um conjunto de dados meteorológicos, de várias estações distribuídas, em vários municípios do Paraná, Santa Catarina, São Paulo e Mato Grosso do Sul (Figura 3.1), tendo como principal foco a coleta de dados meteorológicos do Estado do Paraná.

Estes dados foram cedidos pela Fundação ABC e abrangem 115 áreas de monitoramento, em um período aproximado de 11 anos (1999 a 2010), de observações. Para esta pesquisa, a amostragem utilizada foi de aproximadamente 60, das 115 áreas de monitoramento, onde se encontram as estações.

Figura 3.1 - Distribuição das estações meteorológicas



Fonte: O autor

A partir dos dados meteorológicos disponibilizados, por meio de arquivos texto, os mesmos foram importados em um banco de dados *DB2 Express-C*, que além de tratar de forma eficiente os dados convencionais (alfanuméricos), possui uma extensão de tratamento de dados espacial (*Spatial Extender*) o qual integrado com seus recursos nativos de programação (*functions* e *store procedures*), permite realizar testes de validação e estimativa.

Quanto à utilização de um SGBD, com suporte a recursos espaciais, se deve ao fato do mesmo facilitar e agilizar a identificação da distância e posição de cada estação meteorológica com base na posição de suas coordenadas, além de permitir o armazenamento das demais informações a serem tratadas.

Estas estações coletam dados, tais como: temperatura do ar, umidade relativa, precipitação, velocidade do vento, radiação solar, dentre outras variáveis climáticas. A periodicidade de coleta desses dados é feita de uma em uma hora.

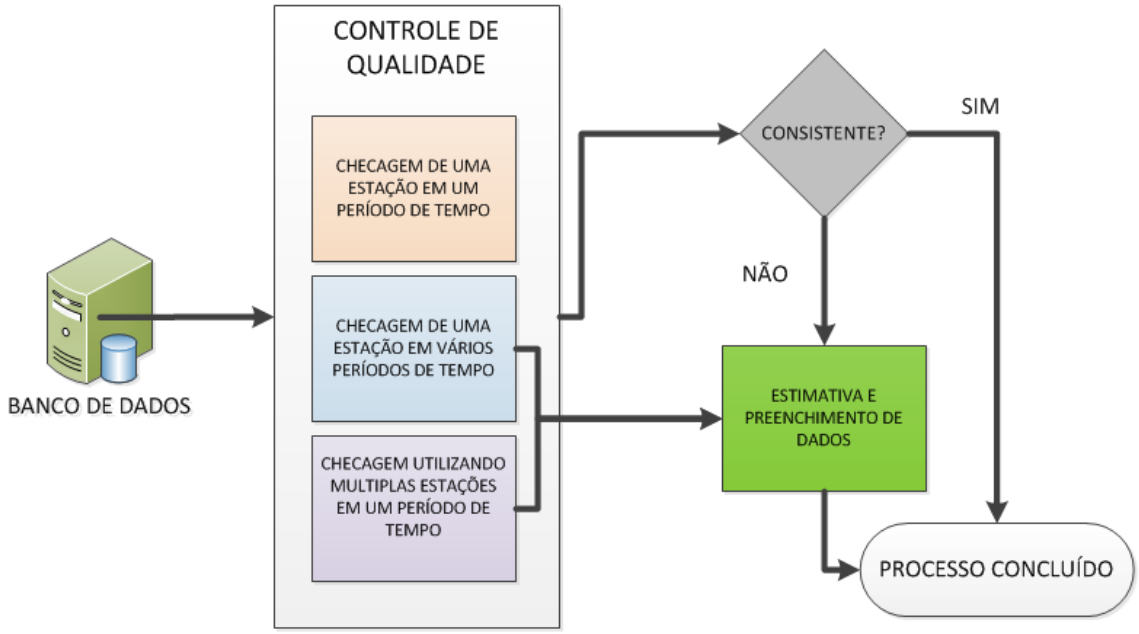
A variável testada nesta pesquisa foi a temperatura, evitando se assim repetições de testes com a mesma técnica.

3.2 METODOLOGIA

A metodologia proposta nesta pesquisa, para a identificação dos erros e inconsistências e falhas de dados, usa de um conjunto de regras de validação, que iniciam com validações simples de restrição de valores, por meio de limites, até

verificações mais complexas com base nas informações geradas, pelas estações vizinhas, no mesmo período de tempo (Figura 3.2).

Figura 3.2 - Visão geral da metodologia



Fonte: O autor

A princípio, as etapas de validação e consistência tornaram-se especialmente importantes visto que, evita ou alerta a entrada de dados incorretos, ajudando na identificação de dados suspeitos, aprimorando assim a qualidade dos mesmos.

Este é um dos princípios básicos para todo e qualquer sistema de informação, onde se busca corrigir ou mesmo bloquear a entrada de dados incorretos e inconsistentes *(“garbage in, garbage out”*).

A validação dos dados segue alguns dos princípios propostos por Mateo & Leung (2010), utilizando-se de estatística. Já para as etapas de estimativas de dados, especialmente aqueles que estão ausentes da série, foram utilizados métodos estatísticos e geoestatísticos, como krigagem e Inverso do Quadrado da Distância (IQD).

Posteriormente, são comparados os dados reais com os gerados pelos métodos estatísticos. A informação estimada deve levar em conta o aspecto espacial e temporal.

3.2.1 Validação e controle de qualidade de dados

O sistema de controle de qualidade e validação de dados, aqui proposto, tem como principal objetivo identificar dados incorretos e, principalmente, dados suspeitos de estarem com problemas. Esta etapa é fundamental para o aprimoramento dos dados armazenados, ficando de fora os dados ausentes visto que os mesmos são facilmente identificáveis.

Este controle está dividido em 3 partes:

a) Verificação de dados da estação, em um período específico do tempo, baseando-se em limites e restrições simples (validação básica);
b) Verificação dos dados da estação, com base no seu histórico de comportamento, ou seja, uma validação com base na série temporal; e
c) Verificação dos dados, com base o comportamento das estações vizinhas, ou seja, uma validação de caráter espacial.

3.2.1.1 Validação básica

Esta etapa verifica a ocorrência de dados obviamente inconsistentes. Geralmente, estes dados são aqueles gerados por erros de leituras dos sensores, podendo estar com problemas de calibração ou com defeitos.

a) *Validação de limites*: aqui é verificado se os valores a serem analisados respeitam algumas condições básicas da variável a qual representa, ou seja, se ela representa um valor que é fisicamente possível de ser obtido. Sob este ponto de vista, pode-se restringir os dados válidos dentro de um limite de valores possíveis. Por exemplo: a umidade relativa do ar deve estar em um intervalo de dados de 0% a 100%, a temperatura que deve estar em um intervalo de -90ºC e +60ºC. Qualquer valor fora deste intervalo está obviamente incorreto, pois representa um valor impossível, para a variável em questão.

b) *Validação lógica*: muitos dos registros das estações meteorológicas registram dados médios, máximos e mínimos, referente ao intervalo de tempo do monitoramento. Estes dados devem respeitar suas condições lógicas básicas. Por exemplo, a temperatura média de um período não pode ser maior que a temperatura máxima ou menor que a temperatura mínima deste mesmo

período. Nesse caso, a Regra $Temp_{minima} \leq Temp_{média} \leq Temp_{máxima}$ deve ser respeitada.

c) *Validação de limites do período:* esse método, ao contrário da validação anterior, busca identificar valores suspeitos de erros. Com base na comparação dos valores de um intervalo específico, verifica se os mesmos apresentam condições físicas realmente passíveis de acontecerem. Por exemplo, caso a coleta de dados de uma estação, com periodicidade horária que fornece valores de temperatura que atendam aos requisitos das validações anteriores, no caso, que esteja dentro de um limite físico permitido e cujas temperaturas mínimas, médias e máximas estejam sendo respeitadas. Ainda assim, deve-se verificar se a diferença entre a temperatura máxima e a temperatura mínima, ocorridas nesta uma hora de medições, respeita as condições físicas possíveis. Nesse exemplo poderia ser definido algo, como a diferença de temperatura máxima permitida para uma hora de medição é de 10ºC, na equação $Temp_{máxima} - Temp_{minima} \leq 10\,℃$. Mesmo assim, qualquer situação que não respeite esta regra não deve ser descartada, mas sim apresentada como um dado suspeito de conter falhas, visto que variações extremas de mudanças de clima são passíveis de acontecer. Nesta pesquisa, além da verificação entre os períodos horários, foi realizada também uma verificação com a diferença de temperatura do dia a qual não deve ultrapassar $25\,℃$.

3.2.1.2 Validação temporal

Após a validação de dados básica, onde as inconsistências básicas foram identificadas, uma nova etapa de validação faz-se necessária, baseada no comportamento histórico da variável, ao longo das medições passadas, baseada em sua série temporal.

As validações básicas apresentadas identificam uma série de problemas, nos valores coletados em uma estação. No entanto, existe a necessidade de verificar se esta informação é consistente para o período, em que foi gerada. Por exemplo, alguns fenômenos relacionados ao clima, geralmente, seguem certos padrões, como os relacionados à temperatura. De uma forma geral, durante os períodos de verão são registradas as temperaturas mais altas do ano, assim como

no inverno existe a tendência de serem registradas as temperaturas mais baixas. Com base nesta premissa faz necessário verificar se o valor gerado pela estação está compatível com os valores para aquele período. Por exemplo, valores próximos de 0ºC (zero) no verão, possivelmente, represente uma informação incorreta, assim como valores muito elevados no inverno, acima de 25ºC ou 30ºC podem indicar problemas nas leituras.

Esta validação não alerta a presença de valores incorretos, mas sim de valores suspeitos de conterem erros, visto que fenômenos atípicos e quebras de recordes históricos podem acontecer.

Para este caso utilizou-se a mesma técnica proposta por Mateo & Leung (2010) para avaliar o aspecto temporal da variável. Esta técnica segue os seguintes passos:

a) Verificação dos valores do dia anterior. Caso a medição seja feita várias vezes no dia, utiliza-se dos valores do dia anterior para o mesmo horário.
b) Verificação dos dados do dia seguinte. Da mesma forma que a verificação anterior, deve-se respeitar a unidade de coleta dos dados.
c) Verificação dos dados referentes aos três dias (anterior, corrente, posterior), dos anos anteriores.

A Tabela 1 apresenta quais são os períodos dos dados que serviram de base, para a validação temporal, com as regras de verificação citadas.

Tabela 3.1 – Exemplo de dados de temperatura média do período da Estação 1

Estação 1			
Ano	**Dia anterior**	**Data analisada**	**Dia posterior**
	01/11	02/11	03/11
2009	21,3	22,1	21,9
2008	23,2	21,1	22,0
2007	22,4	22,1	22,0
2006	21,3	21,1	22,1
2005	22,9	23,2	23,0
...			
1999	22,9	22,2	23,3

Fonte: O autor

Neste exemplo, a data a ser analisada é do dia 02/11/2009, cuja temperatura média é avaliada com base nos valores dos dias 01/11/2009 (dia anterior) e

03/11/2009 (dia posterior), e ainda toda a série histórica do período correspondente a estes três dias, ou seja, o período de 01/11 a 03/11, de 1999 a 2009.

Como se pode perceber esta validação não é realizada no momento em que o dado é inserido, pois ela depende dos dados do dia posterior, para completar o conjunto dos dados para validá-lo. No entanto, caso seja necessário realizar um processamento de validação temporal, no momento em que o dado é inserido, pode-se utilizar apenas o dia anterior e a série histórica do período de três dias.

De posse dos valores históricos da estação, verifica-se então por meio do cálculo do escore Z (equação 3.1), o quanto o valor analisado se comporta, quando comparado com os demais, no mesmo período.

$$Z = \left| \frac{x_d - \bar{x}}{\sigma} \right| \quad (3.1)$$

O escore Z vai avaliar quantas vezes o valor analisado se afastou do desvio padrão, a partir da média da amostragem. Por exemplo, supondo que a média do período seja 20^0C, o desvio padrão seja 2^0C e o valor a ser avaliado seja 24^0C, de acordo com a equação do escore Z, tem-se um resultado de 2 que significa que 24^0C se afastou duas vezes o valor do desvio padrão (2^0C), a partir da média 24^0C. Nesta pesquisa utilizou-se a variável temperatura com um valor aceitável, para o escore Z, de 3. Este valor de escore foi igual ao proposto por Mateo & Leung (2010).

Por meio da validação temporal, foi possível ajustar alguns valores de validação de limites. No caso da temperatura foi elaborada uma restrição de temperatura de -20C a +45ºC. No entanto, esta é uma restrição global e cada local onde as estações estão instaladas possuem suas próprias restrições. Neste caso foi feita uma leitura dos registros, de cada estação, e criada os registros de recordes de temperatura, máxima e mínima, para cada uma delas (Tabela 3.2). Esta informação também servirá no auxílio à verificação dos dados.

Tabela 3.2 - Exemplo de restrição de temperatura por estação

Estação	Temperatura Mínima	Temperatura Máxima
Castro	-3,90	33,00
Ilha do Mel	5,80	37,10
Ibaiti	2,40	34,60
Ivaí	-1,80	34,50
Inácio Martins	-2,30	31,10
Morretes	3,30	41,70

Fonte: O autor

3.2.1.3 Validação espacial

As duas formas de validação, anteriormente descritas, permitem a verificação de inconsistências de dados das estações, de forma individual, com base no conjunto de informações gerado, ou seja, considerando apenas as informações geradas por elas mesmas sem, no entanto, considerar o comportamento das estações que estão próximas.

Nesta técnica são utilizados os dados de várias das estações que se encontram próximas espacialmente no sentido de vizinhança, para detectar erros ou valores suspeitos, uma vez que foi assumido o princípio de que existe uma correlação ou um padrão de comportamento quanto a sua distribuição espacial entre as estações mais próximas.

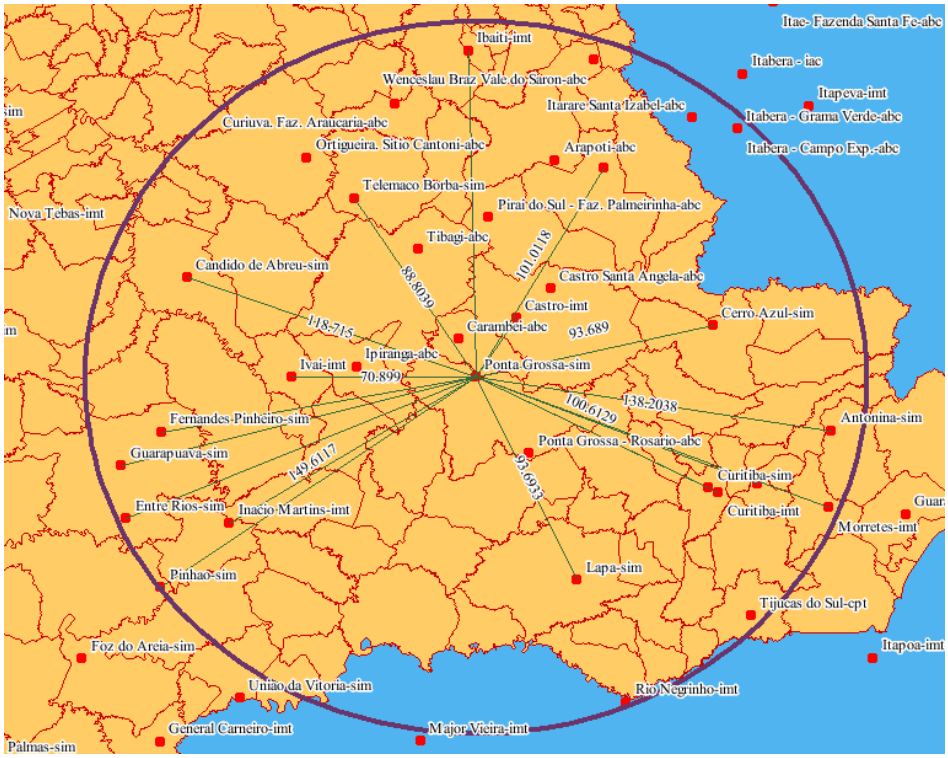
Para fazer esta verificação foram utilizadas técnicas estatísticas, como a correlação de Pearson, para verificar quais as estações estão mais correlacionadas umas com as outras; a técnica de regressão linear, para efetuar as estimativas de cada estação; e de testes de intervalo de confiança, a fim de verificar quais valores são significativamente diferentes, das observações das estações vizinhas.

Cada etapa de aplicação destas técnicas é descrita a seguir:

a) Identificação das estações mais próximas

   Partindo do princípio que estações mais próximas tem maior probabilidade de ter comportamentos semelhantes, com relação às variáveis tratadas, e para evitar o cálculo de correlação de forma desnecessária, entre estações distantes, restringiu-se o raio de abrangência do calculo de correlação. Nesta pesquisa optou-se em utilizar um raio máximo de até 150 km, entre as estações meteorológicas, exemplificado na Figura 3.3.

Figura 3.3 - Exemplo do raio de 150 km a partir da estação de Ponta Grossa



Fonte: O autor

b) Cálculo da correlação de Pearson

Após a identificação das estações próximas utilizou-se do cálculo de correlação para cada par de estações para a variável temperatura.

Para gerar o coeficiente de correlação utilizou-se todos os registros históricos de temperatura, de cada uma das estações, limitando aquelas que possuíam informações de um mesmo período. De posse dos coeficientes de correlação, entre as estações, convencionou-se a Tabela 3.3 para interpretar o grau de correlação entre elas. Esta convenção assemelha-se a proposta de Callegari-Jacques (2003) citada por Lira (2004).

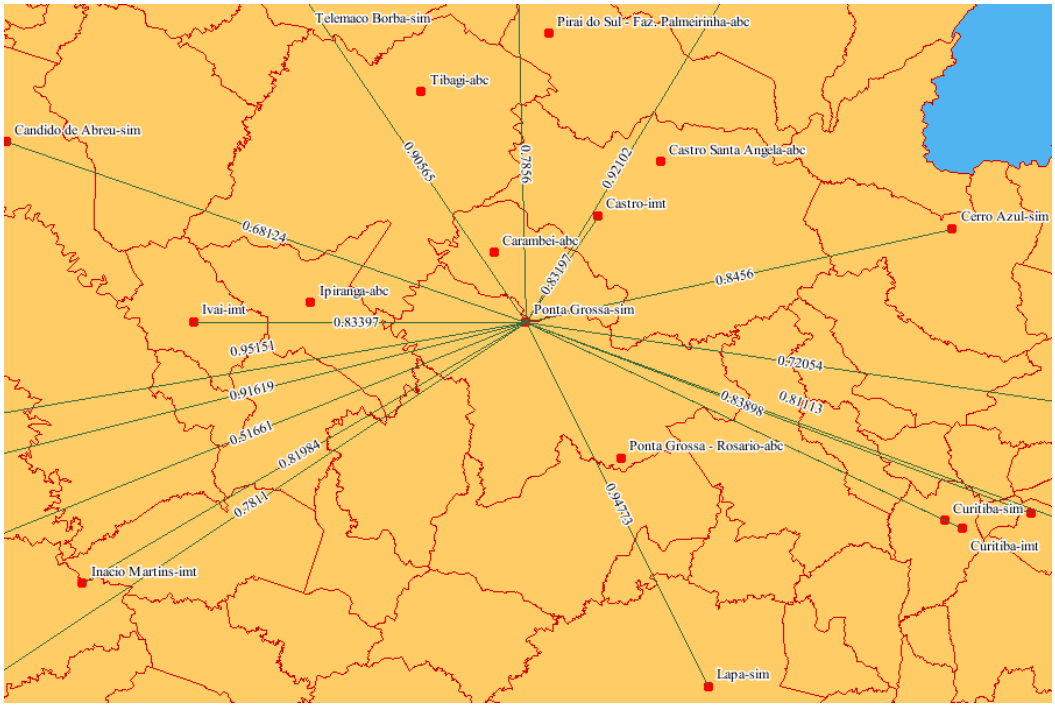
Tabela 3.3 - Interpretação dos índices de correlação

Valor de ρ (+ ou -)	Interpretação
0.00 a 0.19	Correlação bem fraca
0.20 a 0.39	Correlação fraca
0.40 a 0.69	Correlação moderada
0.70 a 0.89	Correlação forte
0.90 a 1.00	Correlação muito forte

Fonte: O autor

Nos cálculos de validação espacial foram consideradas, apenas, as estações cujo grau de correlação fosse igual ou superior a 0,85, indicando assim uma forte correlação entre as estações. Figura 3.4 apresenta os índices de correlação entre algumas estações meteorológicas.

Figura 3.4 - Correlação entre a estação de Ponta Grossa e as estações vizinhas



Fonte: O autor.

c) Estimativa de dados mediante regressão linear

Após selecionar as estações com bom índice de correlação, foi utilizada a correlação linear (equação 3.2) para gerar um valor estimado para a estação a serem validados os dados, com base nos dados das estações próximas.

$$\hat{x}_{i,d} = a_i + b_i\, y_{i,d} \qquad (3.2)$$

O calculo da regressão é feito para cada uma das estações próximas. De forma similar, para cada valor estimado através da regressão linear, calculou-se a raiz do erro médio quadrático, comparado com os valores observados, conforme a equação 3.3.

$$RMSE_i = \sqrt{\frac{\sum_{d=1}^{N}(x_d - \hat{x}_{i,d})^2}{N}} \qquad (3.3)$$

Sendo, x_d é o valor observado;

$\hat{x}_{i,d}$ é o valor estimado pela regressão linear; e

N é o número de observações.

Finalmente, com base em todos os erros médios quadráticos gerados em cada uma das estações próximas, foi gerado um novo valor estimado para a estação sendo avaliada. Esse novo valor é utilizado como variável, para ponderar o dado em questão, conforme equação 3.4.

$$w_d = \frac{\sum_{i=1}^{m} \frac{y_{i,d}}{RMSE_i{}^2}}{\sum_{i=1}^{m} \frac{1}{RMSE_i{}^2}} \quad (3.4)$$

Sendo, m é o número de pares de estações; e

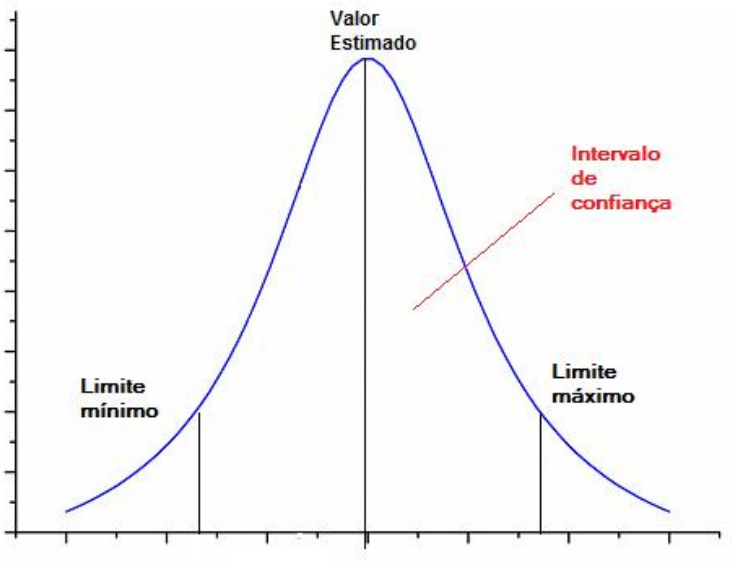
$y_{i,d}$ é o dado observado na estação próxima i, em um período d.

Após estimar o valor calcula-se o erro padrão, conforme equação 3.5:

$$SE = \sqrt{\frac{m}{\sum_{i=1}^{m} \frac{1}{RMSE_i{}^2}}} \quad (3.5)$$

De posse de todos estes dados se pode então verificar se o dado observado se encontra dentro de um limite de dados aceitável o qual caracteriza o intervalo de confiança da informação (Figura 3.5).

Figura 3.5 - Representação de um intervalo de confiança



Fonte: O autor

A equação 3.6, baseada na metodologia de Mateo & Leung (2010), foi utilizada para verificar o intervalo de confiança:

$$w_d - c \times SE \leq x_d \leq w_d + c \times SE \quad (3.6)$$

Dessa forma, dizemos que o valor observado x_d deve estar em um intervalo de valores, baseado no valor estimado, com base em todas as estações próximas w_d. Esse valor pode variar, tanto para menos quanto para mais, baseado no erro padrão.

O valor de c é uma constante, podendo variar conforme a variável analisada. Nesta pesquisa, para a variável temperatura foi assumido o valor 4, sendo este valor um pouco menos restritiva quando comparada a proposta de Mateo e Leung (2010), que utilizou o valor de 3. Esta constante c serve para indicar quantas vezes, além do erro padrão, será o intervalo de confiança tanto para mais, quanto para menos. Qualquer valor que esteja fora deste limite é considerado um dado suspeito e passível de ser analisado.

3.2.2 Estimativa de dados por meio de Interpolação

Nesta pesquisa foram utilizados dois métodos de estimativa de dados, levando em consideração o aspecto espacial. O primeiro baseado em geoestatística utilizando de krigagem, e o segundo, foi o IQD.

Ambos os métodos, utilizam-se da distância entre os pontos, para determinar o valor do ponderador, o qual posteriormente foi utilizado em conjunto com o valor observado nas estações, gerando o valor estimado para um determinado ponto.

No caso da krigagem, foi utilizada a geoestatística para avaliar a variabilidade espacial do atributo temperatura, sendo a dependência espacial expressa por meio do semivariograma, estimado pela equação 3.7.

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(X_i) - Z(X_i + h)]^2 \quad (3.7)$$

Onde, N(h) é o número de pares de valores medidos Z(xi), Z(x+h), separado pela distância h. O gráfico de $\hat{\gamma}(h)$ versus os valores correspondentes de h é chamado de semivariograma.

Para estimar valores de Temperatura, utilizou-se dois métodos de

interpolação, a krigagem ordinária, definido pela Equação 3.8.

$$\hat{Z}(x_i, x_i + h) = \sum_{i=1}^{n} \lambda_i Z(x_i, x_i + h) \quad (3.8)$$

sujeito a:

$$\sum_{i=1}^{n} \lambda_i = 1 \quad (3.9)$$

Em que, $\hat{Z}(x_i, x_i + h)$ é o estimador para um ponto $(x_i, x_i + h)$ da região e λ_i são os pesos usados na estimativa.

O outro método utilizado foi o inverso do quadrado da distância que é um interpolador determinístico univariado, de médias ponderadas. Esse método é definido pela equação 3.10 (MELLO et al., 2003):

$$x_p = \frac{\sum_{i=1}^{n}\left(\frac{1}{d_i^2} * x_i\right)}{\sum_{i=1}^{n}\left(\frac{1}{d_i^2}\right)} \quad (3.10)$$

Onde, x_p = atributo interpolado; x_i = valor do atributo do i-ésimo ponto de amostragem; d = distância euclidiana entre o i-ésimo ponto de vizinhança e o ponto amostrado; e n = número de amostras.

De forma similar, ao que foi feito com a regressão linear, foi estipulado e aplicado um limite máximo de alcance a fim de limitar o raio de atuação para os cálculos. Este limite estipulado em 100 km de raio, ou seja, todas as estações dentro deste limite passam a fazer parte do cálculo da estimativa (Figura 3.6).

Nesses dois métodos, a distância passa a ser parte fundamental do cálculo do estimador e, provavelmente, estações muito distantes pouco contribuíram no cálculo da estimativa. Talvez, poderiam contribuir para gerar distorções. Já no método estatístico anterior (regressão linear), existe a questão do índice de correlação, a qual descartava as estações com baixa correlação, do cálculo da estimativa.

Figura 3.6 - Visualização do raio de alcance de 100 km de uma estação

Fonte: O autor

4 RESULTADOS E DISCUSSÃO

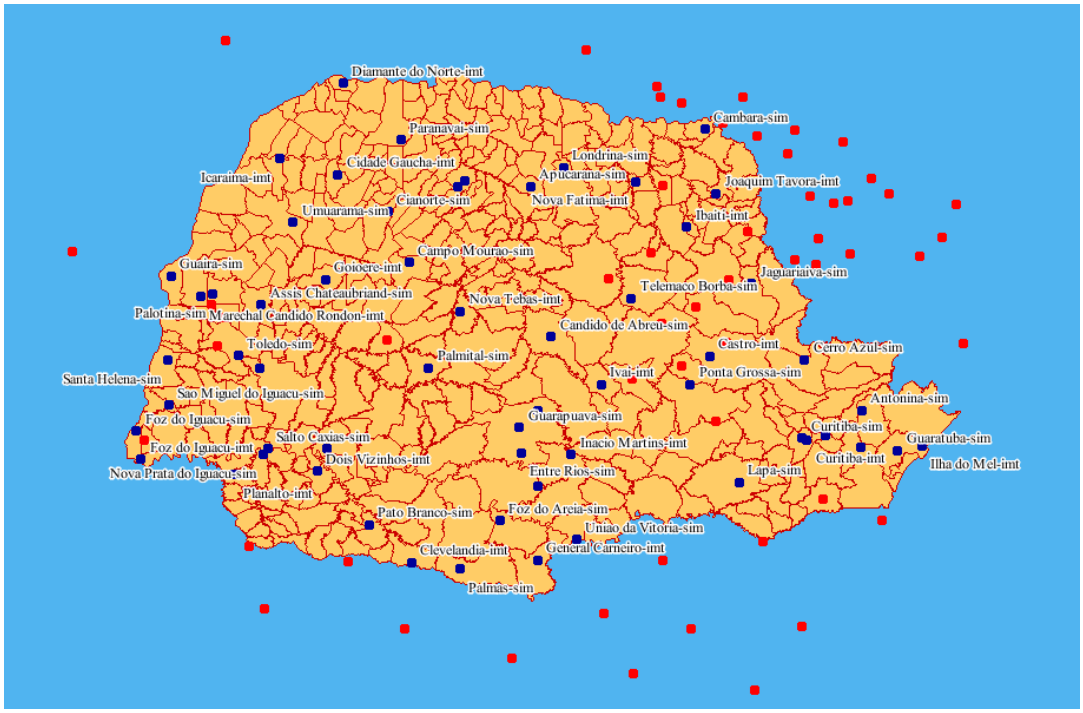
Para analisar o desempenho das técnicas, foram realizados ensaios na identificação de dados inconsistentes e na estimativa de dados. Os ensaios se restringiram a variável temperatura, com as seguintes estações: Curitiba, Ponta Grossa, Entre Rios, Foz do Areia, Guarapuava, Lapa, Palmas, Pinhão e União da Vitória. O período onde foram estimados os valores ficou restrito a novembro de 2009.

4.1 VALIDAÇÃO BÁSICA

Antes da aplicação das estimativas foi realizada uma avaliação de todos os dados da variável temperatura, a fim de verificar os problemas básicos, assim como os dados suspeitos. Esta avaliação não se restringiu as estações e períodos mencionados, pois o intuito era ter uma visão geral dos problemas, mais comuns, de erros ou dados suspeitos.

O total de dados importados aproximou-se a 4 milhões de registros, correspondendo aos dados de 1999 a 2010, de 60 estações (Figura 4.1).

Figura 4.1 - Visualização das estações com dados fornecidos no estudo



Fonte: O autor

De todo universo de dados, foi possível detectar problemas tanto de ordem básica, quanto aqueles considerados suspeitos. Seguindo a metodologia proposta, seguem os resultados de cada uma das etapas.

A Tabela 4.1 apresenta as restrições aplicadas e a quantidade de registros encontrados que não atendiam aos critérios exigidos.

Com base nos resultados, foi possível verificar a ocorrência de dados, cuja qualidade está comprometida por não atender aos requisitos mínimos de qualidade impostos pela validação básica.

Tabela 4.1 - Validações básicas dos dados

Regra de validação (Temperatura)	**Quantidade de Registros ~= 4 milhões**	**Percentual %**
Validação de Limites -10℃ < Temp < 45℃	14.459	0,35%
Validação Lógica TempMin < TempMed < TempMax	24.746	0,60%
Período sem informação	37.787	0,92%
Validação de Limites por período de 1 hora superior a 10℃. (TempMax – TempMin) > 10℃	4.436	0,11%
Validação de Limites por período de 1 dia superior a 25℃. (TempMax – TempMin) > 25℃	1.718	0,04%

Fonte: O autor

Na Figura 4.2 é apresentado um totalizador dos dados inconsistentes, sendo que este totalizou 1,9% da amostragem, evidenciando assim a importância da verificação dos dados e a necessidade de um tratamento aprimorado. Entretanto vale ressaltar que este percentual pode aumentar, significativamente, caso sejam feitas verificações semelhantes nas demais variáveis coletadas pelas estações não contempladas (pressão, umidade relativa do ar, etc.).

Figura 4.2 – Resultado geral das validações

Fonte: O autor

Para analisar o desempenho da técnica de validação básica, foram realizados ensaios na identificação de inconsistências e erros de dados.

Foram realizados os testes de validação básica em cada uma das estações a fim de verificar possíveis erros e inconsistências, no entanto, os testes ficaram restritos as estações e períodos de interesse, os resultados são apresentados na Tabela 4.2.

Tabela 4.2 - Resultado da validação básica nas estações novembro de 2009

Tipo de verificação/ Estação	União da Vitoria	Curitiba	Lapa	Guarapuava	Foz do Areia	Entre Rios	Pinhão	Ponta Grossa	Palmas
Validação de Limites -10℃ < Temp < 45℃	0	0	0	0	1	0	0	0	0
Validação Lógica TempMin < TempMed < TempMax	0	0	0	0	0	0	0	0	0
Período sem informação	5	0	0	1	0	1	0	0	0
Validação de Limites por período de 1 hora superior a 10℃. (TempMax – TempMin) > 10℃	0	0	0	0	1	7	1	0	0
Validação de Limites por período de 1 dia superior a 25℃. (TempMax – TempMin) > 25℃	0	0	0	0	1	2	0	0	0
Total	**5**	**0**	**0**	**1**	**3**	**10**	**1**	**0**	**0**

Fonte: O autor

Pode-se observar que mesmo em um intervalo curto de tempo (um mês), para um número restrito de estações (nove), foi possível observar erros de dados, quanto à ausência de informações. No caso dos dados com erro de valores, foi identificada a Estação Foz do Areia, onde houve ocorrência de temperatura máxima de 74,5℃, evidenciando o erro na informação. Nenhuma ocorrência de erros de lógica foi identificada em nenhum das estações. No entanto, foram detectadas algumas falhas de dados, em especial, na Estação de União da Vitória que apresentou cinco períodos sem dados.

Também foi possível identificar a ocorrência de dados com suspeita de erros. A primeira com verificação de diferenças de temperaturas no período de uma hora, sendo detectando três estações. A Estação de Pinhão apresentou uma ocorrência, com uma diferença de temperatura horária de 11,2℃, ou seja, 1,2℃ acima do limite variação de 10℃ por hora. No caso da Estação Entre Rios, foi identificado o maior número de ocorrências (7), com diferenças de temperaturas de 16,6℃ à 28,7℃.

Na Estação de Foz do Areia, a variação foi mais acentuada, de 51,7 ℃, mas essa variação é consequência do problema identificado anteriormente devido a ocorrência de um valor máximo de temperatura de 74,5℃.

Na análise de variação diária de temperatura (último item da tabela 4.2), até um limite de 25℃, o registro de Foz do Areia (1), que chegou a uma variação diária de 56,4℃, continua sendo consequência do valor distorcido de temperatura máxima de 74,5℃. Já as duas ocorrências de Entre Rios, também, é consequência da validação anterior, de limite horário de variação, estando entre 28,4℃ e 29,4℃.

Vale ressaltar, que um mesmo dado incorreto pode gerar uma sequência de alertas de erro durante cada uma das validações, confirmando assim a importância de verificações simples e básicas.

4.2 VALIDAÇÃO TEMPORAL

Após esta análise inicial da ocorrência de erros, a etapa seguinte envolve a análise temporal dos dados, onde se verifica a ocorrência de dados com suspeitas de erro baseada na sua série histórica, em um determinado período.

A utilização de um critério, no caso o escore Z, permite identificar registros que se distanciaram do comportamento normal daquela região, tendo como base a

sua série histórica, com valores distantes do normal para a região, podendo evidenciar possíveis inconsistências de dados.

Nos testes, adotou-se o critério limite de três (3), para o escore Z, o mesmo adotado pela metodologia de Mateo e Leung (2010). Qualquer dado acima deste limite foi analisado, mas não descartado ou encarado como erro, visto que podem ocorrer casos de quebras de recordes históricos, para um determinado período e local.

Os resultados desta verificação, nas estações em estudo, são apresentados na Tabela 4.3.

Tabela 4.3 - Resultados Escore Z

Estação	Escore Z	Nr. Ocorrências
União da Vitória	4	1
Curitiba	-	-
Lapa	-	-
Guarapuava	-	-
Foz do Areia	4	1
Entre Rios	-	-
Pinhão	4	5
	5	3
	6	1
	7	1
Ponta Grossa	-	-
Palmas	-	-

Fonte: O autor

Pode-se observar, no caso do escore Z, ocorrências de desvios de valores com base na sua série histórica, com destaque para a Estação de Pinhão, onde se identificou os valores com os maiores desvios (acima de 4 e chegando a 7). Verificando este caso, em específico, constatou-se que o valor coletado pela estação no período, com suspeita de problemas, foi de 20,6℃, com um histórico de valores próximos de 27℃.

4.3 VALIDAÇÃO ESPACIAL

A etapa seguinte envolve tanto a validação, quanto a estimativa de valores para as estações. Em um primeiro momento foi gerado o valor estimado com base nas estações com maior correlação de Pearson. Em seguida, o valor estimado serve de base para o cálculo dos valores de limites de tolerância, para que seja comparado com o dado observado, criando uma forma de validação, através de um intervalo de confiança. A quantidade de dados suspeitos registrados pela validação espacial é apresentada na Tabela 4.4.

Tabela 4.4 - Registros suspeitos de erros

Estação	Num. de registros suspeitos	Percentual de registros suspeitos
União da Vitória	25	3,5%
Curitiba	83	11,5%
Lapa	66	9,2%
Guarapuava	71	9,9%
Foz do Areia	46	6,4%
Entre Rios	104	14,4%
Pinhão	67	9,3%
Ponta Grossa	59	8,2%
Palmas	103	14,3%

Fonte: O autor

Mesmo utilizando critérios menos rigorosos para a elaboração do intervalo de confiança (4), o método apresentou uma taxa relativamente alta de registros considerados suspeitos. A razão para esta taxa, certamente, está relacionada com o método regressão linear, o qual em determinadas estações ou determinados horários, apresentou forte tendência em subestimar ou superestimar valores, sendo que este valor estimado pela regressão linear o valor de referência para a geração do intervalo de confiança utilizada na validação espacial.

4.4 ESTIMATIVA DE VALORES

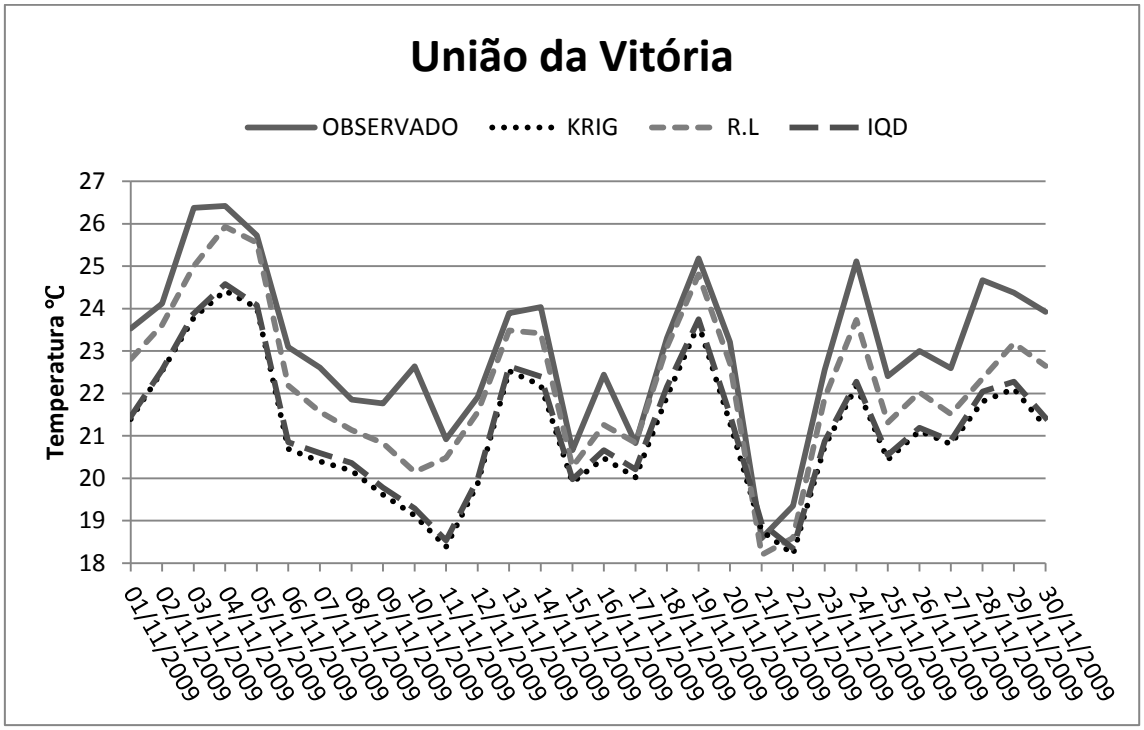
Após a validação, baseada no intervalo de confiança, foi realizada a análise comparativa dos resultados obtidos pelas estimativas de krigagem, inverso do quadrado da distância e regressão linear. Após a geração de cada um dos valores,

foi possível compará-los com os dados observados, sendo necessário agrupar os dados gerados, utilizando a média diária das temperaturas.

Os resultados das comparações são apresentados nas Figuras 4.3 à 4.11, onde é observado o desempenho de cada técnica, para cada uma das estações analisadas.

Para a estação de União da Vitória todas as técnicas apresentaram tendência em subestimar os valores conforme observado na Figura 4.3. Os métodos da krigagem e IQD foram os que apresentaram maior tendência em subestimar os dados (92,6% e 91,2% respectivamente), já o método da regressão linear apresentou uma tendência menor em subestimar os dados (76,6%) além de apresentar valores de Erro Médio (EM) e Raiz do Erro Quadrático Médio (REQM) menores (0,8330 e 1,0021 respectivamente) quando comparados aos da krigagem e do IQD. Esta foi a única das estações cujo número de estações de apoio foram de igual número tanto para a técnica de regressão linear (R.L) quanto para as técnicas de krigagem e IQD.

Figura 4.3 - Resultados da estação de União da Vitória

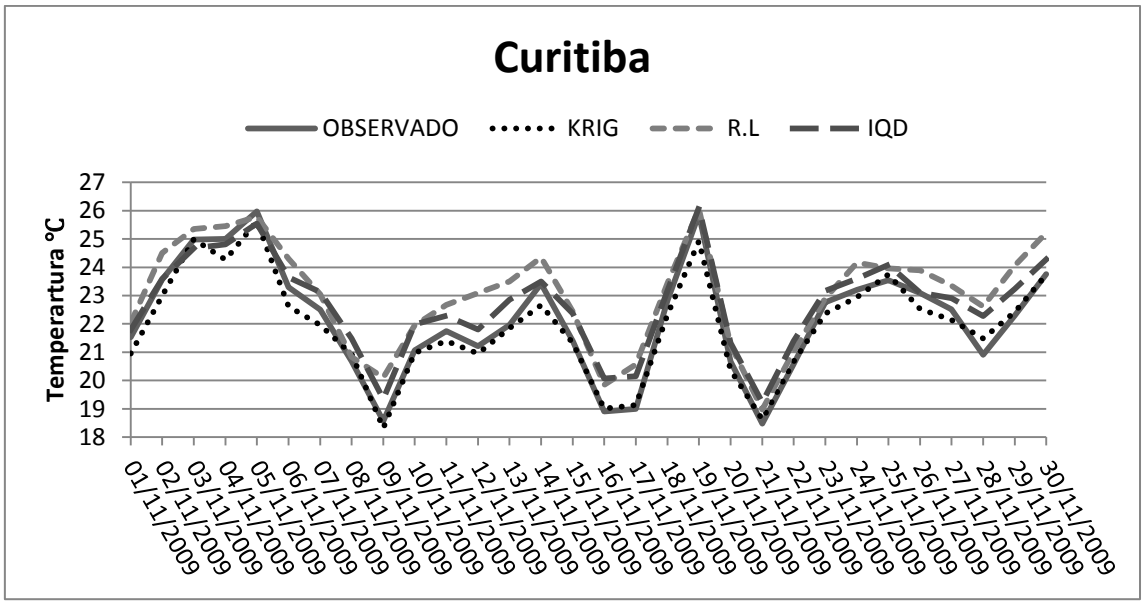


Fonte: O autor

Para a estação de Curitiba destaca-se a técnica da krigagem, pois apresentou EM (0,2367) e REQM (0,4237) menores quando comparados as técnicas

de regressão linear e IQD. No entanto, o método da krigagem foi a que apresentou as maiores diferenças entre os valores observados e os estimados (6,8145℃ e -7,3458℃), já o método da regressão linear foi a que apresentou as menores diferenças (1,365℃ e -4,3624℃). Observa-se pela Figura 4.4 que as 3 técnicas apresentaram valores próximos dos observados.

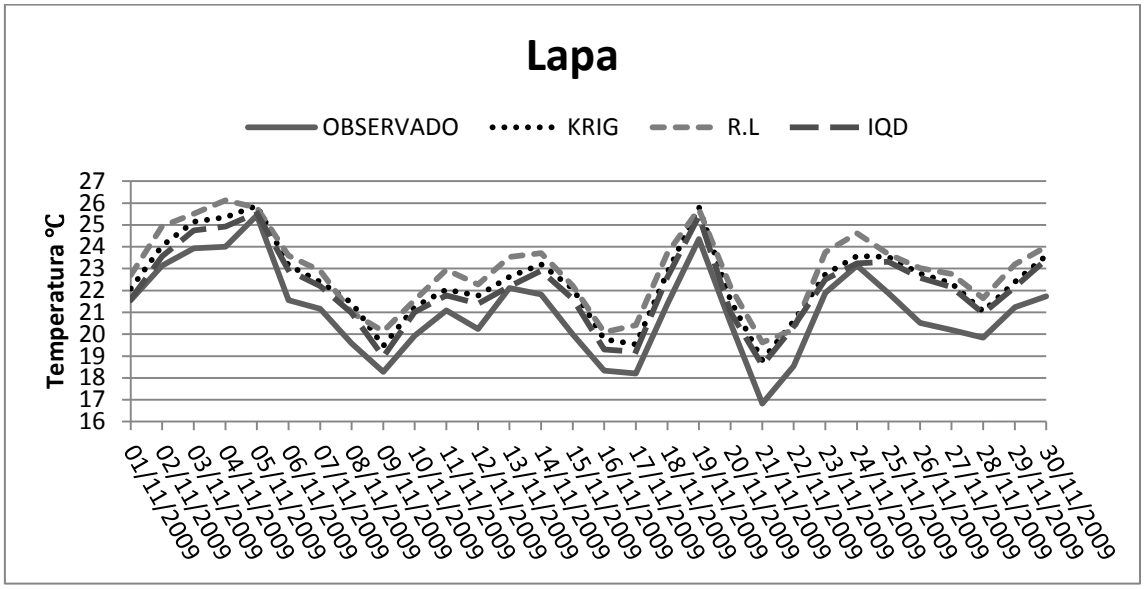
Figura 4.4 - Resultados da estação Curitiba



Fonte: O autor

Para a estação da Lapa observa-se que as 3 técnicas aplicadas apresentaram tendência em superestimar os valores conforme apresentado na Figura 4.5. Para esta estação a técnica da regressão linear foi a que apresentou a maior tendência em superestimar os dados (94%) e também foi a que apresentou os maiores EM e REQM (-1,8402 e 1,8973 respectivamente).

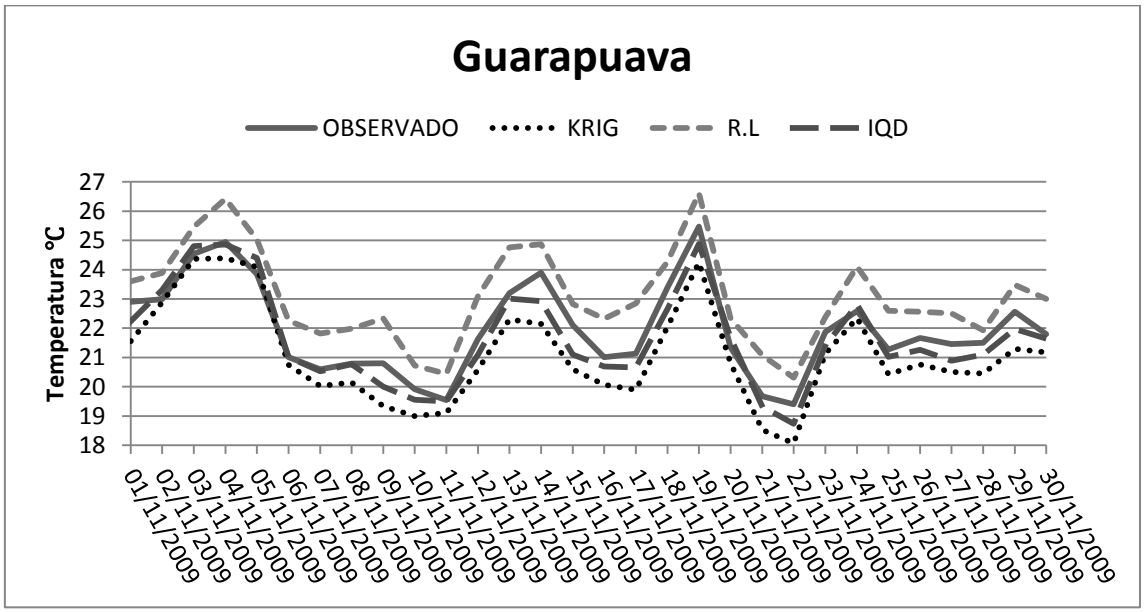
Figura 4.5 - Resultados da estação Lapa



Fonte: O autor

Na estação de Guarapuava o método de regressão linear apresentou tendência em superestimar os dados (88%), enquanto as técnicas de krigagem e IQD apresentaram tendência a subestimar os dados (83% e 66% respectivamente), conforme observado na Figura 4.6. Para esta estação o IQD foi o que apresentou os menores EM e REQM (0,3032 e 0,4919 respectivamente).

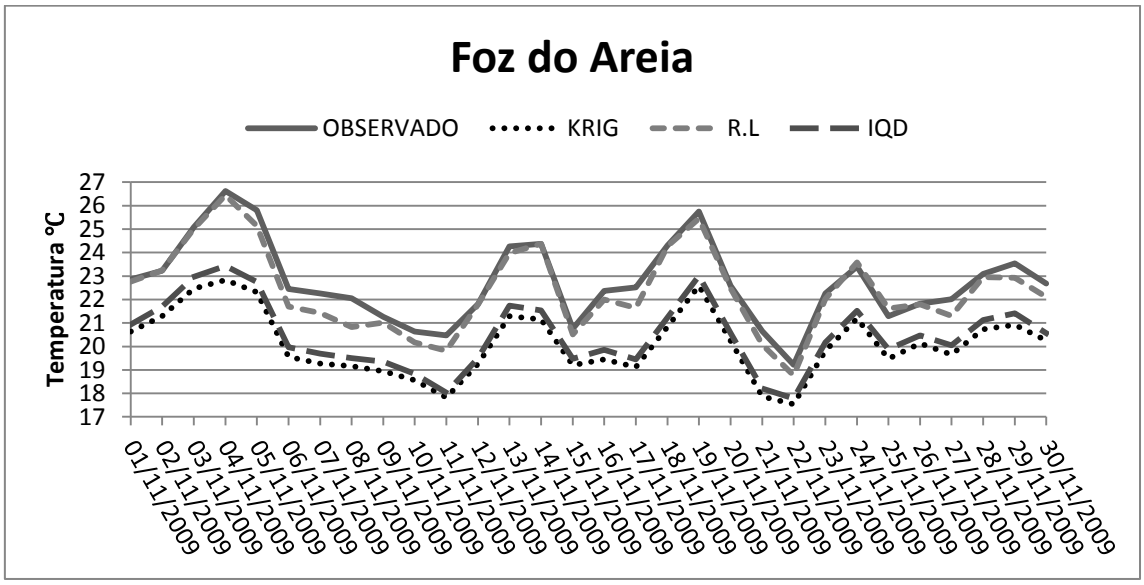
Figura 4.6 - Resultados da estação Guarapuava



Fonte: O autor

Na estação Foz do Areia nota-se que as técnicas da krigagem e IQD subestimaram os dados em mais de 90%. Ambas as técnicas também apresentaram EM (acima de 2) e REQM (acima de 2) semelhantes. Já a técnica da regressão linear apresentou valores de EM e REQM significativamente mais baixos (0,3382 e 0,4838 respectivamente) quando comparados com a krigagem e o IQD.

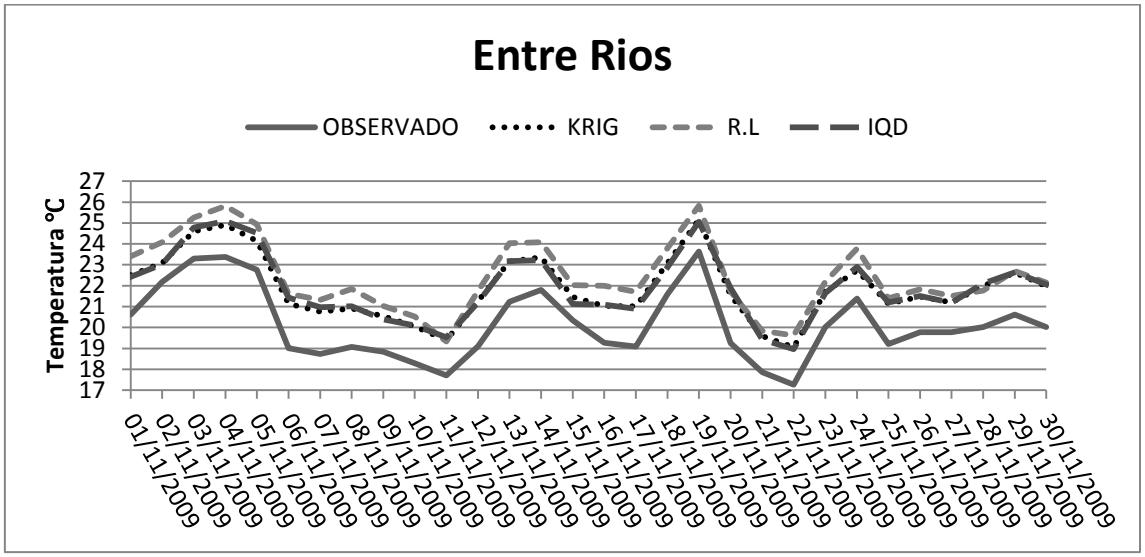
Figura 4.7 - Resultados da estação Foz do Areia



Fonte: O autor

Para a estação de Entre Rios todas as técnicas apresentaram forte tendência em superestimar os valores, conforme observado na Figura 4.8. Destaca-se para esta estação a técnica da regressão linear a qual superestimou em 99,6% das vezes os dados estimados. Foi esta técnica que apresentou as maiores EM e REQM (-2,2594 e 2,2840 respectivamente). Já as técnicas da krigagem e IQD apresentaram valores semelhantes tanto de EM (aproximadamente 1,7) quanto de REQM (aproximadamente 1,7).

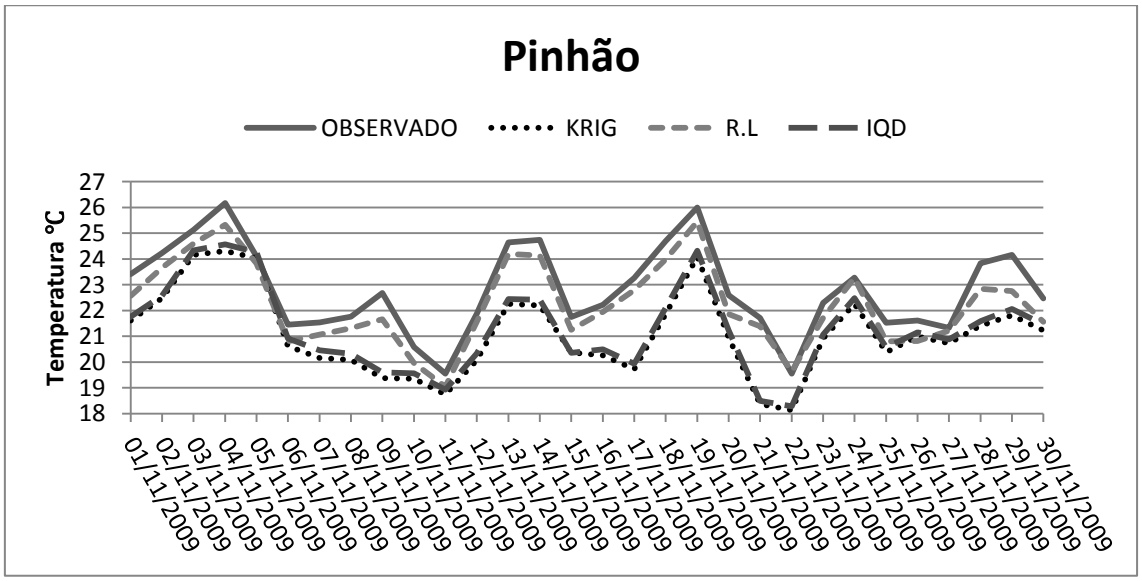
Figura 4.8 - Resultados da estação Entre Rios



Fonte: O autor

Para a estação de Pinhão a técnica de regressão linear apresentou valores de EM e REQM de 0,5721 e 0,6485 estimando assim valores mais próximos dos observados (Figura 4.9) quando comparados as técnicas de krigagem e IQD.

Figura 4.9 - Resultados da estação Pinhão

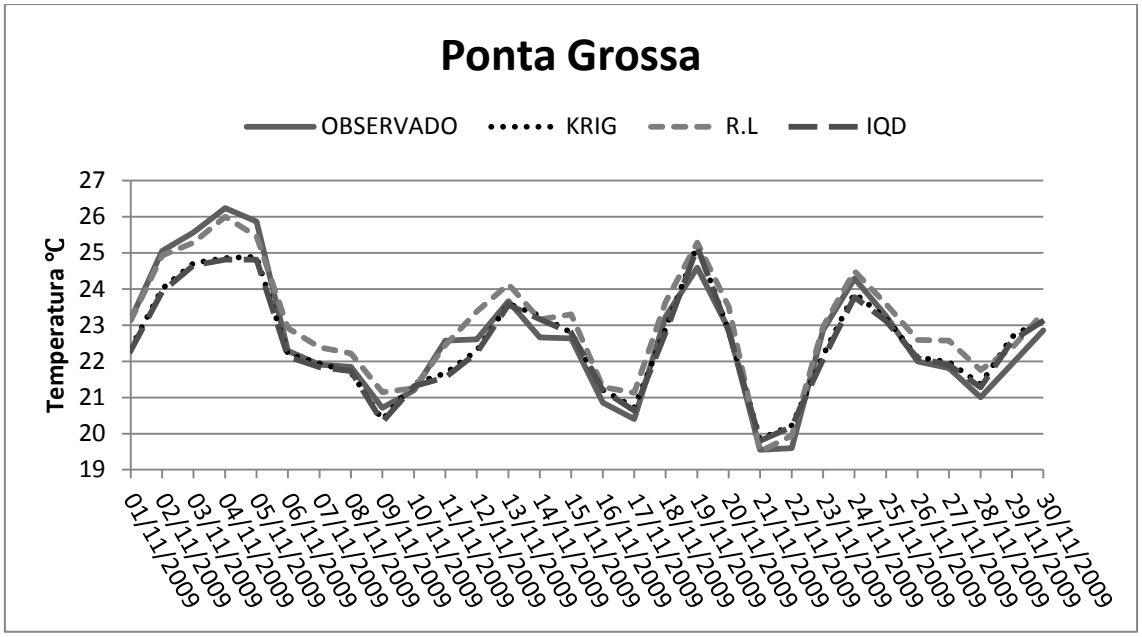


Fonte: O autor

Para a estação de Ponta Grossa observa-se na Figura 4.10 que todas as técnicas apresentaram valores próximos dos observados. Esta estação foi a que utilizou o maior número de estações de apoio (vizinhança), 13 para a krigagem e

IQD e 12 para a regressão linear, de forma que, todas as técnicas apresentaram EM e REQM mais baixos quando comparadas as demais estações.

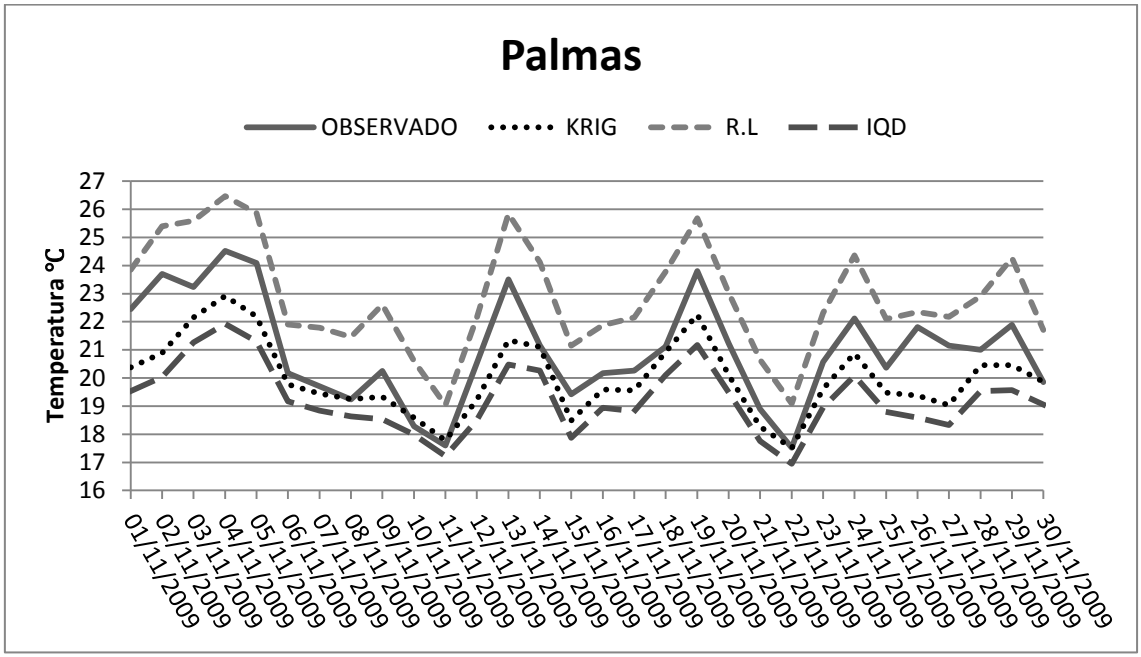
Figura 4.10 - Resultados da estação Ponta Grossa



Fonte: O autor

Na estação de Palmas observa-se que a técnica de regressão linear apresentou forte tendência em superestimar os dados (98%) ocorrendo o oposto (subestimar os dados) nas técnicas de krigagem e IQD (70% e 74% respectivamente). Para o EM e REQM a técnica da krigagem apresentou os menores valores (0,9805 e 1,2781 respectivamente).

Figura 4.11 - Resultados da estação Palmas



Fonte: O autor

Foi possível identificar, por meio dos gráficos, que uma mesma técnica comporta-se de forma diferente, nas diferentes estações. Algumas vezes, com tendências em superestimar quase que totalmente os dados, e outras em subestimar os dados. Por exemplo, na estação de Palmas o método estatístico utilizando regressão linear superestimou quase todos os dados quando comparados com os observados. Já na estação de Pinhão a tendência da mesma técnica foi subestimar os valores.

Foi possível perceber, também, que na maior parte das estimativas, as tendências de comportamento dos dados observados foram seguidas.

Na análise destes dados, através do método de krigagem (Tabela 4.5), constatou-se que poucas vezes houve uma distribuição equilibrada entre dados subestimados e dados superestimados. No entanto, foram observadas algumas exceções, como no caso das Estações de Curitiba (43,8% e 56,3%) e de Ponta Grossa (51,7% e 48,3%), onde houve uma melhor distribuição entre os dados superestimados e os subestimados. Ambas as estações, apresentaram os melhores resultados com relação ao EM e a REQM, Ponta Grossa com 0,1183 e 0,5640 e Curitiba com 0,2367 e 0,4237 respectivamente. Estas foram as que mais tiveram estações de apoio no cálculo das estimativas (13 estações), evidenciando a importância da quantidade de amostras no cálculo das estimativas por krigagem. Já as diferenças máximas de temperaturas comparadas aos dados observados foram obtidas pela estação de Guarapuava onde o valor máximo subestimado foi de 4,5714℃ e o maior valor superestimado foi de 3,38℃.

Tabela 4.5 - Resultados da Krigagem

Estação	Nr. Estações	EM	REQM	Max. Diferença (+)	Max. Diferença (-)	% Pos.	% Neg.
União da Vitória	9	1,9111	2,0380	8,0413	-2,6193	92,6%	7,4%
Curitiba	13	0,2367	0,4237	6,8145	-7,3458	43,8%	56,3%
Lapa	9	-1,3479	1,4379	2,9117	-8,5227	10,4%	89,6%
Guarapuava	10	0,8615	0,9780	4,5714	-3,3800	83,2%	16,8%
Foz do Areia	9	2,5946	2,6555	6,6746	-2,7280	93,9%	6,1%
Entre Rios	9	-1,7157	1,7430	2,2411	-8,4027	7,0%	93,0%
Pinhão	9	1,7072	1,9009	6,3061	-8,0665	86,4%	13,6%
Ponta Grossa	13	0,1183	0,5640	5,9325	-6,4065	51,7%	48,3%
Palmas	10	0,9805	1,2781	6,7813	-6,8108	70,0%	30,0%
Media		0,5940	1,4466	5,5861	-6,0314	59,9%	40,1%
Max		2,5946	2,6555	8,0413	-2,6193		
Min		-1,7157	0,4237	2,2411	-8,5227		

Fonte: O autor

Os resultados obtidos pelo IQD (Tabela 4.6) foram semelhantes aos da krigagem, visto que ambos são baseados na distância, para gerar o ponderador, utilizado no cálculo da estimativa. De forma similar, este método obteve valores de EM e REQM mais baixos, nas estações com maior quantidade de dados de referência vizinha. Destaque neste resultado, para a estação de Guarapuava que apresentou um desempenho semelhante e algumas vezes até melhores, comparada a estações com maior número de amostras com relação ao EM (0,3032) e REQM (0,4919).

Outra questão foi a distribuição dos dados estimados. Dentre os métodos testados este foi o que obteve uma melhor distribuição média entre dados superestimados (42,4%) e subestimados (57,6%), ainda que a sua tendência seja subestimar.

Tabela 4.6 - Resultados do Inverso do Quadrado da Distância

Estação	Nr. Estações	EM	REQM	Max. Diferença (+)	Max. Diferença (-)	% Pos.	% Neg.
União da Vitória	9	1,76978	1,90456	7,7288	-2,8916	91,2%	8,8%
Curitiba	13	-0,5199	0,6762	5,3955	-6,6871	29,7%	70,3%
Lapa	9	-1,0272	1,1657	3,4609	-8,2101	16,4%	83,6%
Guarapuava	10	0,3032	0,4919	4,2194	-4,4112	66,8%	33,2%
Foz do Areia	9	2,2221	2,2839	5,9163	-2,9505	91,8%	8,2%
Entre Rios	9	-1,7436	1,7860	4,0003	-9,2656	12,5%	87,5%
Pinhão	9	1,5056	1,7226	6,1111	-8,3431	83,5%	16,5%
Ponta Grossa	13	0,1823	0,5852	6,4765	-6,5673	52,8%	47,2%
Palmas	10	1,7311	1,9504	7,7871	-6,3137	74,2%	25,8%
Media		0,4915	1,3963	5,6773	-6,1822	57,6%	42,4%
Max		2,2221	2,2839	7,7871	-2,8916		
Min		-1,7436	0,4919	3,4609	-9,2656		

Fonte: O autor

A aplicação do método de regressão linear (Tabela 4.7) apresentou, de forma geral, uma tendência a superestimar os valores, entretanto o comportamento deste método foi semelhante ao apresentado pela krigagem e pela aplicação do IQD, onde os melhores resultados foram obtidos nos locais com maior amostragem.

Este método foi o que apresentou a melhor média de REQM (1,2083) e, também, que apresentou as menores variações entre o valor máximo e mínimo dos erros, superestimando em média 2,4266℃ e subestimando em média 5,140℃. Apesar, deste método, utilizar em média, um número menor de amostragem, no caso as estações de apoio (9,4 contra 10,1 no caso da krigagem e do IQD), apresentou resultados semelhantes de erros e estimativas, quando comparados aos da krigagem e do IQD.

Tabela 4.7 - Resultados do Método de Regressão Linear

Estação	Nr. Estações	EM	REQM	Max. Diferença (+)	Max. Diferença (-)	% Pos.	% Neg.
União da Vitória	9	0,8330	1,0021	5,4797	-2,0016	76,6%	23,4%
Curitiba	8	-0,8477	0,9976	1,3650	-4,3624	30,0%	70,0%
Lapa	7	-1,8402	1,8973	0,7734	-5,7712	5,3%	94,7%
Guarapuava	11	-1,0984	1,1429	2,7359	-5,4700	11,5%	88,5%
Foz do Areia	10	0,3382	0,4838	2,9225	-2,7744	66,7%	33,3%
Entre Rios	10	-2,2594	2,2840	0,3233	-9,6309	0,4%	99,6%
Pinhão	10	0,5721	0,6485	2,8944	-3,4730	71,1%	28,9%
Ponta Grossa	12	-0,3334	0,4741	2,4627	-3,8630	43,9%	56,1%
Palmas	7	-1,8879	1,9445	2,8827	-8,9137	1,7%	98,3%
Media		-0,7249	1,2083	2,4266	-5,1400	34,1%	65,9%
Max		0,8330	2,2840	5,4797	-2,0016		
Min		-2,2594	0,4741	0,3233	-9,6309		

Fonte: O autor

Um aspecto analisado foram os dados estimados, localizados nas bordas de todo o conjunto de estações. Com base em sua localização, em relação às demais estações, consideraram-se como “bordas” as Estações da Lapa, União da Vitória e Palmas (Figura 4.12). Entretanto, não foi constatado que estas estações tivessem o desempenho dos valores estimados, muito diferentes das estações que não estavam nas bordas, visto que, geralmente estações que se encontram nas bordas apresentam resultados menos eficientes de dados estimados, quando comparados

com as estações que não se encontram nas bordas. Provavelmente, esta diferença de resultados não ficou muito evidente nesta pesquisa, devido a quantidade restrita de estações de apoio, tanto as estações que estavam nas bordas, quanto as que não estavam.

Figura 4.12- Visualização de algumas das estações consideradas como "bordas"

Fonte: O autor

Um comparativo geral é apresentado nas Tabelas 4.8 à 4.12, evidenciando em quais pontos cada uma das técnicas apresentou melhor resultado.

A Tabela 4.8 apresenta os resultados comparativos das técnicas sob o aspecto do erro médio absoluto, onde o método de regressão linear apresentou o melhor resultado (1,1274) comparado as demais técnicas. Já a krigagem e o IQD apresentaram valores similares de erro, com uma diferença de aproximadamente 0,05. Quando comparadas ao do método de regressão linear esta diferença passa a ser de 0,20 e 0,15 respectivamente.

Analisando a questão dos erros médios, constata-se que o inverso da distância apresentou o melhor resultado médio de erro (0,4915) comparado com as demais técnicas. Já o método de regressão linear apresentou o menor erro médio máximo (0,833) e a krigagem o menor erro médio mínimo (-1,7157)

Tabela 4.8 - Comparativo dos Resultados (EMA e EM)

			Erro Médio Absoluto			Erro Médio		
Estação	Nr. Estações (100 KM)	Nr. Estações (150 KM)	Krig.	R.L.	IQD	Krig.	R.L.	IQD
União da Vitória	9	10	1,9224	0,8335	1,7926	1,9111	0,8330	1,7698
Curitiba	13	8	0,3454	0,8607	0,5824	0,2367	-0,8477	-0,5199
Lapa	9	7	1,3479	1,8402	1,0272	-1,3479	-1,8402	-1,0272
Guarapuava	10	11	0,8778	1,0984	0,4145	0,8615	-1,0984	0,3032
Foz do Areia	9	10	2,5946	0,3727	2,2221	2,5946	0,3382	2,2221
Entre Rios	9	10	1,7157	2,2594	1,7436	-1,7157	-2,2594	-1,7436
Pinhão	9	10	1,7072	0,5778	1,5166	1,7072	0,5721	1,5056
Ponta Grossa	13	12	0,4473	0,4165	0,4566	0,1183	-0,3334	0,1823
Palmas	10	7	1,0140	1,8879	1,7311	0,9805	-1,8879	1,7311
		Média	1,3302	1,1274	1,2763	0,5940	-0,7249	0,4915
		Máximo	2,5946	2,2594	2,2221	2,5946	0,8330	2,2221
		Mínimo	0,3454	0,3727	0,4145	-1,7157	-2,2594	-1,7436

Fonte: O autor

Quanto ao erro médio percentual (Tabela 4.9), percebe-se que erro médio do IQD e da krigagem apresentam valores próximos de erro 2% e 2,4% respectivamente, enquanto o método de regressão linear apresentou um erro de -3,6%. Percebe-se ainda que no caso da krigagem os erros médios nas localidades com maior quantidade de estações (Curitiba e Ponta Grossa, cada uma com 13 estações) apresentaram desempenho bem superior as demais localidades com menos estações de apoio, com erros de 0,4% à 1%. O método IQD também apresentou resultados melhores nestas localidades, pois tanto quanto a krigagem,

utiliza as mesmas estações como base para as estimativas.

Analisando a REQM o melhor desempenho fica por conta do método de regressão linear que apresentou o resultado médio de 1,2083, ante 1,4466 e 1,3963 da krigagem e do IQD respectivamente.

Observa-se que existe uma alternância do melhor desempenho para cada uma destas análises, representando assim um equilíbrio de desempenho entre as técnicas (Tabela 4.9).

Tabela 4.9 - Comparativo dos Resultados (EMP e REQM)

			Erro Médio percentual			Raiz do Erro Quadrático Médio		
Estação	Nr. Estações (100 KM)	Nr. Estações (150 KM)	Krig.	R.L.	IQD	Krig.	R.L.	IQD
União da Vitória	9	10	8,2%	3,6%	7,6%	2,0380	1,0021	1,9046
Curitiba	13	8	1,0%	-3,9%	-2,5%	0,4237	0,9976	0,6762
Lapa	9	7	-6,6%	-8,9%	-5,0%	1,4379	1,8973	1,1657
Guarapuava	10	11	4,0%	-5,0%	1,4%	0,9780	1,1429	0,4919
Foz do Areia	9	10	11,4%	1,5%	9,7%	2,6555	0,4838	2,2839
Entre Rios	9	10	-8,6%	-11,3%	-8,8%	1,7430	2,2840	1,7860
Pinhão	9	10	7,4%	2,5%	6,6%	1,9009	0,6485	1,7226
Ponta Grossa	13	12	0,4%	-1,5%	0,7%	0,5640	0,4741	0,5852
Palmas	10	7	4,4%	-9,0%	8,0%	1,2781	1,9445	1,9504
		Média	2,4%	-3,6%	2,0%	1,4466	1,2083	1,3963
		Máximo	11,4%	3,6%	9,7%	2,6555	2,2840	2,2839
		Mínimo	-8,6%	-11,3%	-8,8%	0,4237	0,4741	0,4919

Fonte: O autor

Os resultados da Tabela 4.10 apresentam as diferenças máximas tanto positivas (subestimativa) quanto negativas (superestimativa) dos valores estimados em comparação com os valores observados. No caso dos valores máximos positivos o método de regressão linear foi o que apresentou os melhores resultados em todas as estações (média de 2,43), para o valor máximo negativo também apresentou o melhor resultado médio (-5,14). Sob este aspecto o método de regressão linear se sobressaiu comparado as demais técnicas.

Tabela 4.10 – Comparativo de Resultados (Max. diferença positiva e negativa)

			Max. Diferença (+)			Max. Diferença (-)		
Estação	Nr. Estações (100 KM)	Nr. Estações (150 KM)	Krig.	R.L	IQD	Krig.	R.L.	IQD
União da Vitória	9	10	8,041	5,480	7,729	-2,619	-2,002	-2,892
Curitiba	13	8	6,815	1,365	5,396	-7,346	-4,362	-6,687
Lapa	9	7	2,912	0,773	3,461	-8,523	-5,771	-8,210
Guarapuava	10	11	4,571	2,736	4,219	-3,380	-5,470	-4,411
Foz do Areia	9	10	6,675	2,923	5,916	-2,728	-2,774	-2,951
Entre Rios	9	10	2,241	0,323	4,000	-8,403	-9,631	-9,266
Pinhão	9	10	6,306	2,894	6,111	-8,067	-3,473	-8,343
Ponta Grossa	13	12	5,933	2,463	6,477	-6,407	-3,863	-6,567
Palmas	10	7	6,781	2,883	7,787	-6,811	-8,914	-6,314
		Média	5,59	2,43	5,68	-6,03	-5,14	-6,18
		Máximo	8,04	5,48	7,79	-2,62	-2,00	-2,89
		Mínimo	2,24	0,32	3,46	-8,52	-9,63	-9,27

Fonte: O autor

O comparativo da Tabela 4.11 mostra com qual frequência cada técnica superestimou ou subestimou os valores observados. Percebe-se que em muitos casos os métodos apresentaram sempre uma tendência na estimativa, ou frequentemente subestimando ou superestimando os valores. A estação que apresentou a melhor distribuição foi a de Ponta Grossa, aproximando-se de um valor de 50%. Como já mencionado, a estação de Ponta Grossa é uma das que apresentam maior número de estações de apoio, possivelmente esta seja uma das razões do seu melhor desempenho. Já no caso apresentado pelo método de regressão linear para a estação de Entre Rios, mostra uma forte tendência da técnica em superestimar os valores (99,6%), neste caso, esta tendência também foi observada nas outras técnicas. Dentre as técnicas avaliadas a que apresentou a melhor distribuição média de valores foi o IQD com valores de 57,6% e 42,4%.

Tabela 4.11 - Comparativo de Resultados (% Positivos e % Negativos)

			Krig		R.L.		IQD	
Estação	Nr. Estações (100 KM)	Nr. Estações (150 KM)	% Pos.	% Neg.	% Pos.	% Neg.	% Pos.	% Neg.
União da Vitória	9	10	92,6%	7,4%	76,6%	23,4%	91,2%	8,8%
Curitiba	13	8	43,8%	56,3%	30,0%	70,0%	29,7%	70,3%
Lapa	9	7	10,4%	89,6%	5,3%	94,7%	16,4%	83,6%
Guarapuava	10	11	83,2%	16,8%	11,5%	88,5%	66,8%	33,2%
Foz do Areia	9	10	93,9%	6,1%	66,7%	33,3%	91,8%	8,2%
Entre Rios	9	10	7,0%	93,0%	0,4%	99,6%	12,5%	87,5%
Pinhão	9	10	86,4%	13,6%	71,1%	28,9%	83,5%	16,5%
Ponta Grossa	13	12	51,7%	48,3%	43,9%	56,1%	52,8%	47,2%
Palmas	10	7	70,0%	30,0%	1,7%	98,3%	74,2%	25,8%
		Média	59,9%	40,1%	34,1%	65,9%	57,6%	42,4%
		Máximo	93,9%	93,0%	76,6%	99,6%	91,8%	87,5%
		Mínimo	7,0%	6,1%	0,4%	23,4%	12,5%	8,2%

Fonte: O autor

Na análise apresentada pela Tabela 4.12 mostra o desvio padrão dos dados estimados, neste caso, os menores valores médios foram da técnica IQD (3,078). Para a correlação o método de regressão linear apresentou os melhores resultados com índice médio de correlação acima de 0,9. O destaque desta análise fica por conta da estação de Palmas que para o método estatístico obteve valores satisfatórios (0,922) mesmo utilizando menos estações nas estimativas (7). Enquanto as técnicas da krigagem e do IQD, ambas apresentaram graus de correlação mais baixos (0,726 e 0,679) utilizando de 10 estações.

Tabela 4.12 - Comparativo de Resultados (Desvio padrão e Correlação)

			Desvio padrão			Correlação		
Estação	Nr. Estações (100 KM)	Nr. Estações (150 KM)	Krig.	R.L	IQD	Krig.	R.L	IQD
União da Vitória	9	10	3,0070	3,4987	3,0415	0,9048	0,9325	0,9051
Curitiba	13	8	2,7487	3,2384	2,6614	0,8089	0,8948	0,8562
Lapa	9	7	2,9756	3,4046	2,8916	0,9072	0,9096	0,9131
Guarapuava	10	11	3,3105	3,5215	3,0737	0,9499	0,9636	0,9453
Foz do Areia	9	10	3,1775	3,6583	3,1584	0,8563	0,9594	0,8872
Entre Rios	9	10	3,1456	3,6140	3,0621	0,9302	0,9413	0,9056
Pinhão	9	10	2,9874	3,5667	3,0193	0,9111	0,9494	0,9078
Ponta Grossa	13	12	3,0580	3,4645	3,0466	0,8388	0,9438	0,8153
Palmas	10	7	3,5891	3,5024	3,7487	0,7266	0,9226	0,6792
		Média	3,1111	3,4966	3,0781	0,8704	0,9352	0,8683
		Máximo	3,5891	3,6583	3,7487	0,9499	0,9636	0,9453
		Mínimo	2,7487	3,2384	2,6614	0,7266	0,8948	0,6792

Fonte: O autor

Constatou-se um equilíbrio no desempenho dos métodos, cada uma apresentou um resultado melhor em determinada análise. No entanto, foi possível verificar que a quantidade de amostras influi diretamente no desempenho de todas as técnicas.

4.5 TRABALHOS RELACIONADOS

Tsukahara et al (2010) utilizou métodos de redes neurais para a estimativa e preenchimento de dados, permitindo comparação com os resultados deste trabalho. Ambos, trataram dados das mesmas áreas e estações meteorológicas. Para a variável temperatura, os erros médios gerados pelas redes neurais apresentaram valores semelhantes aos deste trabalho. Com relação ao erro quadrático médio, os valores da rede neural foram da ordem de 1,27℃, enquanto que as estimativas de IQD, regressão linear e de krigagem apresentadas neste trabalho apresentaram respectivamente os seguintes resultados: 1,40℃, 1,21℃ e 1,44℃.

Ventura (2012), que também utilizou redes neurais, para estimar dados climáticos, os resultados obtidos para temperatura obteve um erro médio absoluto entre 0,57℃ e 0,77℃. Valores com desempenho mais fracos foram obtidos pelas técnicas de krigagem, regressão linear e IQD aqui apresentados. Os valores foram 1,33℃, 1,12℃ e 1,27℃, respectivamente, destacando a técnica de regressão linear que apresentou melhor resultado.

Apesar do método baseado em regressão linear ter apresentado o menor erro quadrático médio, não significa que ele seja superior ou inferior, aos demais métodos, visto que ficou evidente que a quantidade e a qualidade das estações de apoio têm forte influência no cálculo da estimativa. Desta forma, o desempenho de cada uma destas técnicas pode apresentar resultados diferentes, utilizando-se dados de outras regiões, com densidades diferentes de amostras.

Lado et al. (2007) apresentaram resultados, utilizando-se de métodos estatísticos e geoestatísticos, na estimativa de dados de temperatura. Os bons índices apresentados devem-se ao fato do mesmo utilizar um conjunto de estações bastante grande, o que certamente contribuiu para os bons resultados. Esta questão também é comentada por SOUZA et. al. (2011), onde os métodos como krigagem e inverso da distância não apresentaram bons resultados, na estimativa de dados pluviométricos, sendo apontada como principal causa do fraco desempenho das estimativas a baixa densidade de pontos de coleta de dados (estações).

Nesta pesquisa ficou evidente que o número de dados de apoio tem grande influência na estimativa, visto que os menores erros foram apresentados pelas estações que continham os maiores números de amostras. Outro ponto importante desta pesquisa foi a utilização de um banco de dados espacial, possibilitando a

realização dos cálculos da distância entre os pontos de interesse sem a utilização de outro software de apoio.

Todos os cálculos referentes a krigagem, IQD e regressão linear foram implementados no banco de dados. Esta foi uma característica não encontrada nos trabalhos analisados. Todos os demais trabalhos utilizaram ferramentas estatísticas externas ao sistema de armazenamento de dados, para manipular os dados a serem estimados.

5 CONCLUSÕES E PERSPECTIVAS DE PESQUISAS FUTURAS

No decorrer deste trabalho foi possível verificar que as técnicas empregadas na detecção de erros são pertinentes em estudos meteorológicos com correlação espacial, principalmente, na identificação dos períodos onde os dados apresentavam falhas ou mesmo aqueles que continham inconsistências.

Foi observado que mesmo com uma amostragem pequena, tanto de estações (nove) quanto de período (um mês), é possível identificar ocorrências de erros nos dados, evidenciando assim a frequência com que as falhas e erros ocorrem na coleta de dados. Possivelmente, caso as mesmas técnicas fossem aplicadas as demais variáveis coletadas pelas estações, o número de ocorrência de erros seria ainda maior.

Quanto à validação temporal foi adequada na identificação de dados, onde os valores se distanciavam da média histórica apresentada em um mesmo período. Esta validação é necessária para analisar um grande volume de dados, permitindo identificar de forma eficiente possíveis erros e dados suspeitos.

No entanto, a análise de um especialista da área é necessária, especialmente para dados suspeitos, a fim de verificar se o valor, apesar de diferente do que normalmente é observado, é legítimo, visto que recordes históricos podem ocorrer ou mesmo fenômenos climáticos atípicos.

Quanto às três formas de estimativa de dados, todas apresentaram desempenhos semelhantes, tanto na comparação dos resultados quanto comparadas a trabalhos, com propostas semelhantes de estimativa de dados climáticos. Entretanto, vale ressaltar que o desempenho destas técnicas está diretamente relacionado com a quantidade e a qualidade dos dados utilizados nas estimativas. Certamente, um número maior de estações contribuiria para um melhor desempenho de todas as técnicas analisadas, uma vez que todas as estimativas que utilizaram um maior número de estações apresentaram erros menores.

Dentre as técnicas avaliadas, a regressão linear apresentou resultados satisfatórios, especialmente, pelo fato de ter utilizado em média um número menor de estações para fazer as estimativas. No entanto, quando utilizada esta técnica para gerar o intervalo de confiança para validar os dados observados, esta apresentou comportamento muito restritivo (validação espacial), ou seja, gerou intervalos de confiança que por várias vezes indicavam a existência de dados

suspeitos de erros, quando esta suspeita não se justificava. Na verdade, este foi um problema apresentado quando a técnica superestimava ou subestimava demasiadamente os valores, criando assim intervalos de confiança fora da faixa real dos dados observados.

Outro aspecto observado durante a análise da correlação dos dados das estações foi o fato de que, nem sempre estações mais próximas umas das outras apresentam melhores índices de correlação. Este comportamento ocorreu em vários dos pontos analisados. Avaliando melhor a razão deste comportamento foi possível verificar que estas estações que estavam próximas e apresentavam índices de correlação mais baixos, quando comparadas com outras distantes, foi que a origem dos dados das estações era de instituições diferentes. Isto pode representar que cada instituição tenha seus próprios critérios de calibração dos sensores das estações.

Outro aspecto é que, diferentemente das técnicas de krigagem e do inverso da distância, a regressão linear não permite estimar dados de pontos desconhecidos, visto que ela não usa o aspecto espacial nas estimativas. Já as técnicas de krigagem e inverso da distância permitem que de um determinado ponto (latitude e longitude) seja possível a estimativa de valores, utilizando as estações mais próximas.

É importante ressaltar que o desempenho de cada uma destas técnicas pode apresentar variações, se aplicadas em regiões diferentes e sob um número diferente de amostras. A princípio a estimativa por regressão linear, parece ser a mais indicada quando a quantidade de dados de apoio é pequena, já os métodos IQD e krigagem parecem ser mais adequados quando se têm uma quantidade maior de amostras para a estimativa dos dados.

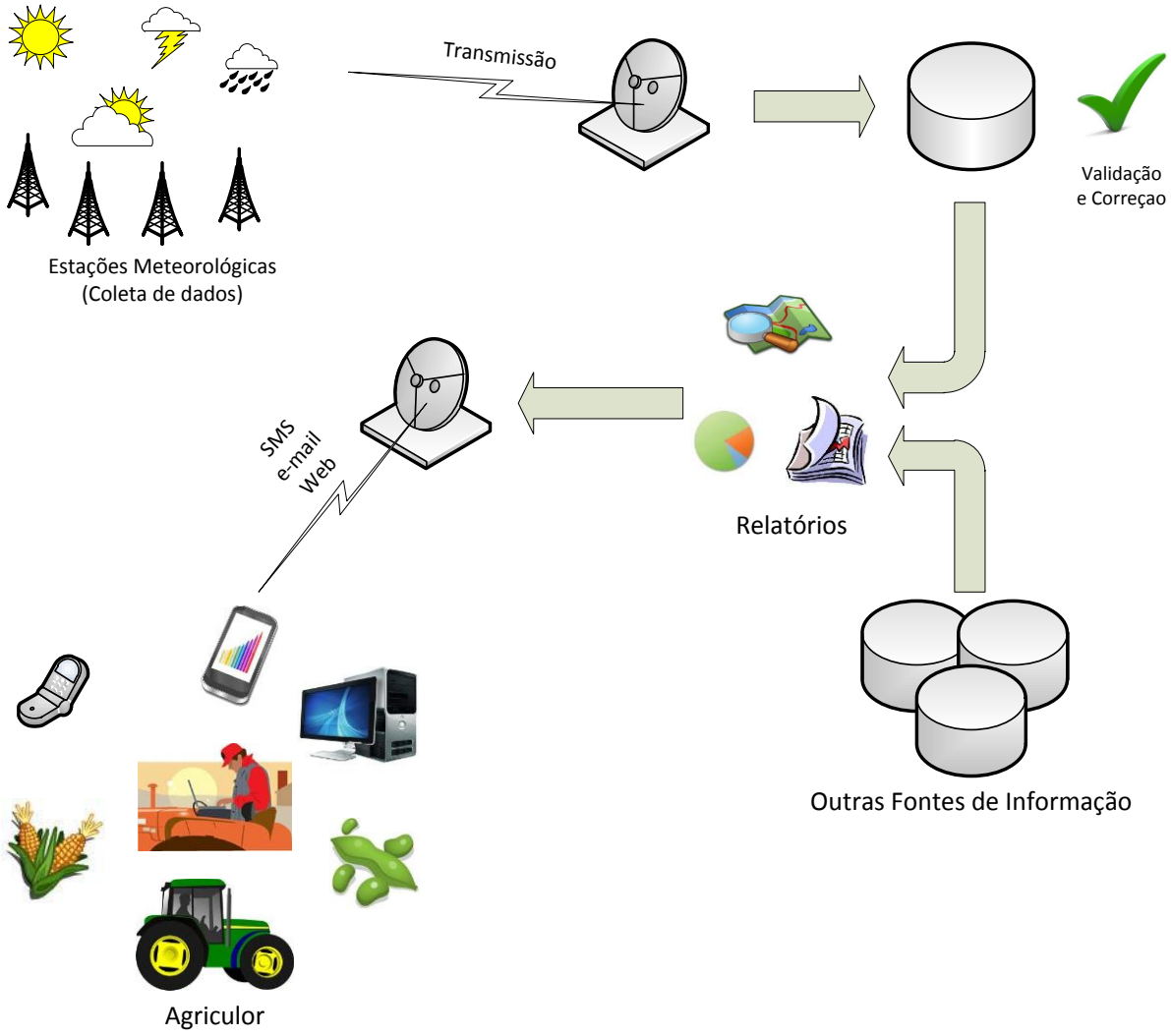
A utilização exclusiva do banco de dados para a geração de todos os testes e dados estimados foi um aspecto relevante. O banco de dados, com capacidade de processamento espacial, é comprovadamente ferramenta eficiente, tanto para o armazenamento dos dados espaciais quanto para seu processamento, permitindo assim a integração tanto dos dados espaciais quanto dos dados convencionais. Por meio dele torna-se desnecessário a utilização de ferramentas externas para efetuar os cálculos, especialmente os de distância, visto que ele já possui todas as funcionalidades.

Esta capacidade permite que os sistemas de validação de dados climáticos,

que necessitam tratar aspectos espaciais, podem realizar as validações no banco de dados. Este aspecto permite ainda que sistemas ou relatórios de alertas de doenças possam ser processados exclusivamente no banco de dados sem a necessidade da utilização de ferramentas ou sistemas externos de processamento de dados espaciais.

Um exemplo seria um sistema de alerta de doenças, onde identificado determinadas condições climáticas em uma região onde uma cultura específica esteja sendo cultivada, permita o envio de um alerta (e-mail, SMS, etc.) a todos os produtores que estão naquela região em um raio de 60 km do foco da doença.

Figura 5.1 - Exemplo de sistema de suporte ao agricultor



Fonte: O autor

Outro ponto que pode ser aprimorado é a combinação de variáveis das próprias estações na validação dos dados ou na estimativa dos mesmos, verificando a correlação entre elas e utilizando esta informação como critério de validação. A utilização de outras técnicas, como co-krigagem e regressão múltipla, pode ser avaliada como trabalhos futuros, a fim de verificar a eficiência quando comparadas com as técnicas empregadas nesta pesquisa.

Desta forma, a proposta deste trabalho foi atingida. As técnicas empregadas permitem a identificação de erros nos dados, assim como alerta possíveis erros de leitura das estações. Os dados estimados por elas apresentaram valores satisfatórios, compatíveis com trabalhos similares. As técnicas estatísticas e geoestatísticas podem auxiliar no controle de qualidade dos dados climáticos, especialmente, por estas gerarem um volume grande de dados, que dificilmente podem ser analisados de forma eficiente, sem a ajuda de um sistema computacional.

REFERÊNCIAS BIBLIOGRÁFICAS

AHRENS, C. D. **Meteorology Today.** 8. ed. Nashville: Thomson Learning, 2006. p. 16-17.

ANDRIOTTI, J. S. **Fundamentos de Estatística e Geoestatística**. Unisinos, São Leopoldo, RS, 2009.

BISQUERRA, R; SARRIERA, J. C; MARTÍNEZ, F. **Introdução à estatística**. Porto Alegre: Artmed, 2007.

BOTELHO, M. F., SILVA C. R.; SCHOENINGER E. R.; CENTENO J. A. S. Comparação dos resultados de interpoladores "Vizinho mais próximo" e "Inverso de uma distância" no cálculo de volume a partir de dados do laser scanner. **Anais XII Simpósio Brasileiro de Sensoriamento Remoto**, Goiânia, Brasil, 16-21 abril 2005.

BURROUGH, P., A., MCDONNEL, R., **Principles of Geographical Information Systems for Lands Resources Assessment**. Oxford, Inglaterra. 1998.

BURROUGH, P.A. **Principles of Geographical Information Systems for Land Resources Assessment**. Clarendon Press, Oxford, Inglaterra. 1986.

CALLEGARI-JACQUES, Sidia M. **Bioestatística: princípios e aplicações**. Porto Alegre: Artemed, 2003. 255p

CORREA S. M. B. B. **Probabilidade e Estatistica**. 2. ed. Belo Horizonte: PUC Minas Virtual, 2003, 116 p.

DRUCK S. **A Geoestatística e os Sistemas de Informações Geográficas**. EMBRAPA, DF, Espaço & Geografia, Vol.5, No 1, 2002.

FERNANDES J. A. B. **Krigagem com deriva externa aplicada a avaliação de recursos minerais de calcário e de minério laterítico**. Dissertação de Mestrado. USP, São Paulo, 2009.

FERREIRA, K. R. **Interface para operações espaciais em banco de dados geográficos**. 2003, 102f. Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais. São José dos Campos, 2003.

FIGUEREDO, D. B. F.; SILVA, J. A. J. Desvendando os mistérios do Coeficiente de Correlação de Pearson. **Revista Política Hoje**, Vol. 18, nr 1, 2009.

FRANKE, R. Scattered Data Interpolation: Test of Some Methods**, Mathematics of Computations**, v. 33, n. 157, p. 181-200. 1982.

GUIMARÃES E. C. **Geoestatística Básica e aplicada**. UFU, Uberlândia – MG, 2004.

HARTKAMP, A.D.; DE BEURS, K.; STEIN, A.; WHITE, J.W. **Interpolation Techniques for Climate Variables**. NRG-GIS Series 99-01. Mexico, 1999.

IBM, **DB2 Spatial Extender e Geodetic Data Management Feature: User Guide and Reference.** Disponível em: < ftp://public.dhe.ibm.com/ps/products/db2/info/vr9 /pdf/letter/en_US/db2sbe90.pdf > , Acesso em 01 nov. 2011

ISAAKS, H. E; SRIVASTAVA, R. M. **Applied geostatistics**. New York: Oxford, 1989.

JAKOB A. A. E. A Krigagem como Método de Análise de Dados Demográficos. UNICAMP/NEPO. **XIII Encontro da Associação Brasileira de Estudos Populacionais**. Ouro Preto, Minas Gerais, 2002.

JAKOB A. A. E.; YOUNG A. F. O uso de métodos de interpolação espacial de dados nas análises sociodemográficas. **XV Encontro Nacional de Estudos Populacionais**, ABEP, MG – Brasil, de 18 a 22 de setembro de 2006.

LADO L. R.; SPAROVEK G.; VIDAL P.T.;DOURADO D.; MACÍAS F. V. **Modelagem da Temperatura do ar para o Estado de São Paulo**. USC/USP/ESALQ, SP, Sci. Agric. (Piracicaba, Braz.), v.64, n.5, p.460-467, September/October 2007.

LANDIM P.M.B. Sobre Geoestatística e mapas. **Terræ Didatica**, Unicamp, 2006.

LIRA S. A. **Análise de correlação: Abordagem Teórica e de construção dos coeficientes com aplicações**. Dissertação – UFPR, 2004.

MATEO M. A. F.; LEUNG C. K. **Design and Development of a Prototype System for Detecting Abnormal Weather Observations** , ACM, 2010.

MELLO C.R.; LIMA J. M.; SILVA A. M.; MELLO J. M.; OLIVEIRA M. S. **Krigagem e o inverso do quadrado da distância para a interpolação dos parâmetros da equação de chuvas intensas.** R. Bras. Ci. Solo, 27:925-933, 2003.

MENDONÇA, F; MORESCO, I. O. **Climatologia: noções básicas e climas do Brasil**. São Paulo: Oficina de Textos, 2007. p. 14-15.

MIRANDA, J. I. **Fundamentos de sistemas de informações geográficas**. EMBRAPA Informação Tecnológica, Brasília, 2005.

NETO P. V. **Estatística Descritiva: Conceitos Básicos**. São Paulo, 2004.

NOGUEIRA, J. D. L.; AMARAL, R. F. Comparação entre os métodos de interpolação (Krigagem e Topo to Raster) na elaboração da batimetria na área da folha Touros – RN. **Anais XIX Simpósio Brasileiro de Sensoriamento Remoto**, Natal, Brasil, 25-30 abril 2009, INPE, p. 4117-4123.

OLIVEIRA M. S.; FERREIRA D. F.; BUENO J. S. S.; LIMA P. C.; LIMA R. R.; VEIGA R. D.; ALVES M. C. **Conceitos e Métodos estatísticos**.

ROHLI, R. V; VEGA, A. J. **Climatology.** 2. ed. Canadá: Jones & Bartlett Learning, 2003. p. 3-4.

SOUZA J. L. L. L.; GOMES T. S.; DIAS R. S.; OLIVEIRA G. M. A.; SANTOS R. L. Avaliação de métodos de interpolação aplicados à espacialização das chuvas no território identidade Portal do Sertão / Bahia. **Anais XV Simpósio Brasileiro de Sensoriamento Remoto - SBSR**, INPE, Curitiba, PR, 2011

TRENTIN G.; HELDWEIN A. B.; STRECK L.; MAAS G. F.; RANDONS S. Z.; TRENTIN R. Controle da requeima em batata cv. 'Asterix' como base para modelos de previsão da doença. **Ciência Rural**, Santa Maria, v.39, n.2, p.393-399, ISSN 0103-8478, mar-abr, 2009.

TSUKAHARA, R.; JENSEN, T.; CARAMORI, P. H. Utilização de Redes Neurais Artificiais para Preenchimento de Falhas em Séries Horárias de Dados Meteorológicos. **XVI Congresso Brasileiro de Meteorologia**, Belém, PA, 2010.

VENTURA, T. M. **Preenchimento de falhas de dados micrometeorológicos utilizando técnicas de inteligência artificial**. Dissertação (Dissertação em Física Ambiental) –UFMT, 2012.

VIOLA, M. R.; MELLO, C. R.; PINTO, D. B. F.; MELLO, J. M.; ÁVILA, L. F**. Métodos de interpolação espacial para o mapeamento da precipitação pluvial**. Revista Brasileira de Engenharia Agrícola e Ambiental. v.14, n.9, p.970–978, 2010.