DATA WAREHOUSE PARA ANÁLISE ESTATÍSTICA DOS ERROS DETECTADOS NO SISTEMA DE ASSIMILAÇÃO DE DADOS DO CPTEC/INPE

[1]*Leopoldo Edgardo Messenger Parada*
[2] *Dirceu Luis Herdies*
[3] *Cristiane Ferreira Lacerda*
[4]*Jorge Luiz Lescura*

Resumo A análise dimensional de fatos de interesse empresarial (ex, vendas e marketing) realizada pelos aplicativos Data Warehouse, é muito usada para decisões estratégicas em empresas. Este trabalho mostra o potencial e os resultados obtidos na aplicação dessa ferramenta computacional na área de meteorologia, em particular, na Assimilação de Dados Meteorológicos do CPTEC/INPE.

Abstract Many companies use dimensional analysis of Data Warehouse systems to support strategies and main business decisions (ie, selling and marketing). This work presents the potential of this computational tool and the results obtained by a system implemented in a meteorological area, particularly, in the data assimilation process of CPTEC/INPE.
Palavras-Chave: Assimilação de Dados, Análise Dimensional.

INTRODUÇÃO

A Assimilação de dados é um processo que envolve a observação do estado da atmosfera, gerada por dados colhidos de diferentes instrumentos meteorológicos, e a informação da previsão dos modelos meteorológicos, fornecendo assim, a melhor estimativa possível do estado real da atmosfera. Uma das principais etapas do Sistema de Assimilação consiste no controle de qualidade dos dados, etapa que tem como resultado um relatório das anomalias detectadas nos dados observacionais. A constatação de anomalias nos dados realizada pelo Sistema de Assimilação é de crucial importância na tarefa de assimilação, no entanto, o sistema PSAS não fornece informações das possíveis origens dos erros, nem avalia em termos percentuais a incidência de anomalias em relação ao total das observações.

Este artigo tem o propósito de descrever e apresentar os resultados obtidos por um sistema Data Warehouse (DW), desenvolvido para monitorar o comportamento do processo de observação de dados do Sistema de Assimilação. O enfoque utilizado consiste na modelagem dimensional dos dados para analisar e estabelecer relações entre os dados observados e o contexto onde os mesmos são gerados. De acordo com a modelagem dimensional, o fato de ocorrer anomalias nos dados pode ser analisado através de várias dimensões diferentes. Entre essas dimensões podemos mencionar, por exemplo, a região onde os dados foram originados, o período da observação, o tipo de

[1] Centro de Previsão de Tempo e Estudos Climáticos – CPTEC/INPE
Centro Universitário Salesiano de São Paulo – UNISAL / Campus Lorena
Rodovia Presidente Dutra Km 40, Caixa Postal 01, Cachoeira Paulista, SP, CEP 12 630 000
Fone (12) 31868533, E-mail leopoldo@cptec.inpe.br

[2] Centro de Previsão de Tempo e Estudos Climáticos – CPTEC/INPE

[3] Centro de Previsão de Tempo e Estudos Climáticos – CPTEC/INPE
Centro Universitário Salesiano de São Paulo – UNISAL / Campus Lorena

[4] Centro de Previsão de Tempo e Estudos Climáticos – CPTEC/INPE

instrumento utilizado e a variável atmosférica observada, citando apenas as dimensões básicas da análise. Outro aspecto relevante do sistema desenvolvido, é que ele permite fazer uma avaliação estatística dos erros, e assim obter um resultado real do andamento da tarefa de observação de dados do Sistema de Assimilação.

DADOS E METODOLOGIA

O processo de Assimilação se inicia com a recepção dos dados no formato ODS (Observation Data Stream) – formato utilizado para dados observacionais. Os dados recebidos são convertidos para um formato do tipo texto e armazenados em uma base de dados relacional como aparece mostrado na figura 1. A partir desse ponto, é necessário converter os dados que se encontram armazenados em uma estrutura relacional para uma estrutura dimensional (cubo) inserida no banco de dados Multidimensional. As consultas de interesse para avaliação dos erros detectados pelo sistema de Assimilação são realizadas nesse último banco.

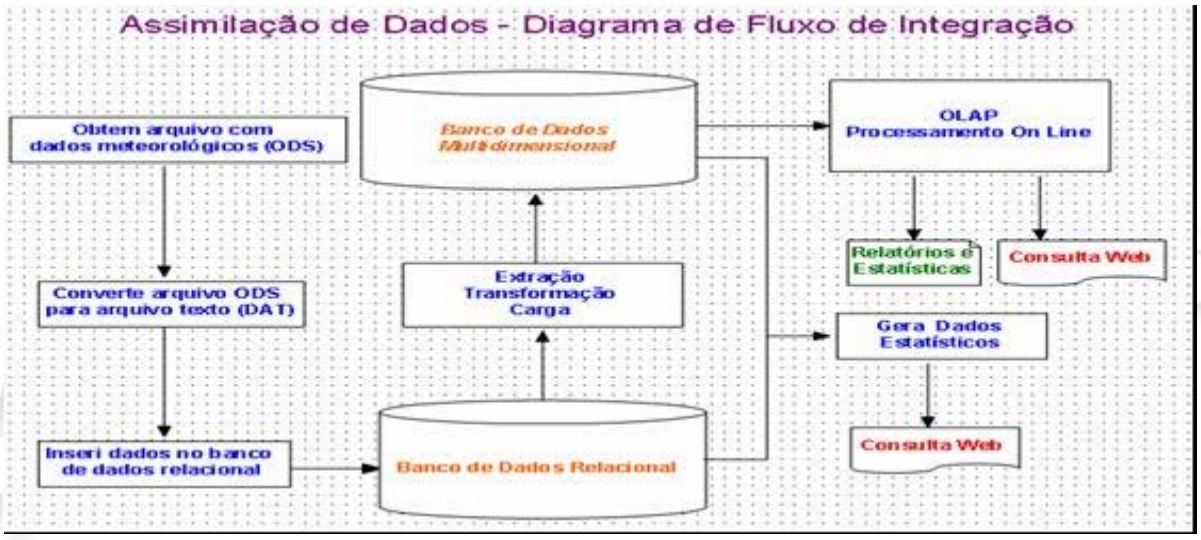

Figura 1. DataWarehouse para análise e estatística de erros na Assimilação de Dados

A análise dimensional de dados utilizada pelo sistema DW pode ser descrita como a observação de fatos de interesse, que ocorrem em um determinado processo. A análise do(s) fato(s) pode ser efetuada a partir de vários pontos de vista simultâneos, que são denominados dimensões. No processo de Assimilação, o *fato* de interesse é a ocorrência de erros (detectados pelo sistema de assimilação de dados) nos dados observacionais recebidos pelo sistema. Não obstante, não existem informações disponíveis que permitam ter alguma idéia de quais são as causas que dão origem a esses erros.

A análise dimensional possibilita observar o fato da ocorrência de erros sob vários aspectos diferentes (dimensões), de modo que, é possível entender melhor a ocorrência dos erros. As dimensões definidas para este sistema são: O tipo de erro, o período de tempo, a localidade, o tipo

de instrumento meteorológico e a variável meteorológica observada.. O sistema implementado pretende verificar a existência, ou não, de relações entre a ocorrência dos erros e alguma(s) dessa(s) dimensões. Por exemplo, os erros podem ocorrer em alguma localidade em particular, ou em algum período de tempo, ou para algum tipo de instrumento meteorológico, ou ainda, para alguma combinação dessas ou outras dimensões.

RESULTADOS

Uma das funcionalidades do sistema DW é a sua capacidade de sumarização de dados. Neste trabalho, foi feita a sumarização dos dados de saída do sistema PSAS (Physical-space Statiscal Analysis System), dados que esse sistema qualifica como dados: *aceitos*, *com anomalias ou rejeitados*. Os dados rejeitados, que são analisados pelo Data Warehouse (DW), correspondem a dados severamente afetados, que não foram utilizados pelo sistema de Assimilação.

O comportamento geral do Sistema de Assimilação pode ser avaliado pela relação: Quantidade Total de Dados Observados versus a Quantidade Total de Dados Rejeitados, como aparece mostrado na figura 2.

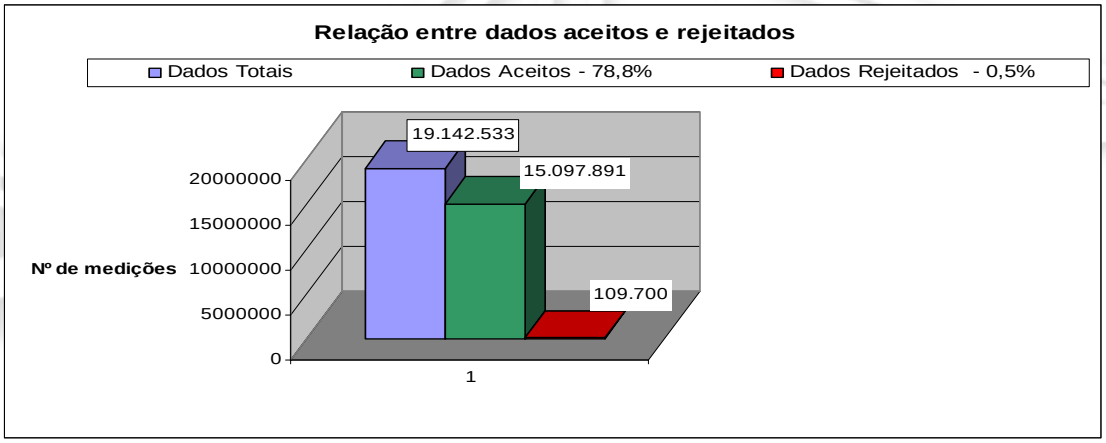

Figura 2. Relação entre dados aceitos e rejeitados – Janeiro 2006

A figura 2 mostra que a porcentagem de dados aceitos é de 78,8% enquanto que a porcentagem de dados com anomalias é de 21,2%. Os dados rejeitados, por sua vez, correspondem a 0,5 % do total de dados.

Além da sumarização de dados, o DW avalia também as possíveis causas que originam os dados rejeitados. Consideram-se as seguintes dimensões de avaliação: período de tempo, instrumento meteorológico, localidade e variável atmosférica.

Avaliação dos erros em termos do período de tempo

A análise, nesta dimensão, é realizada para cada um dos horários sinóticos (00z, 06z, 12z, 18z). Apresentam-se na figura 3 os resultados correspondentes aos meses de janeiro a março de 2006.

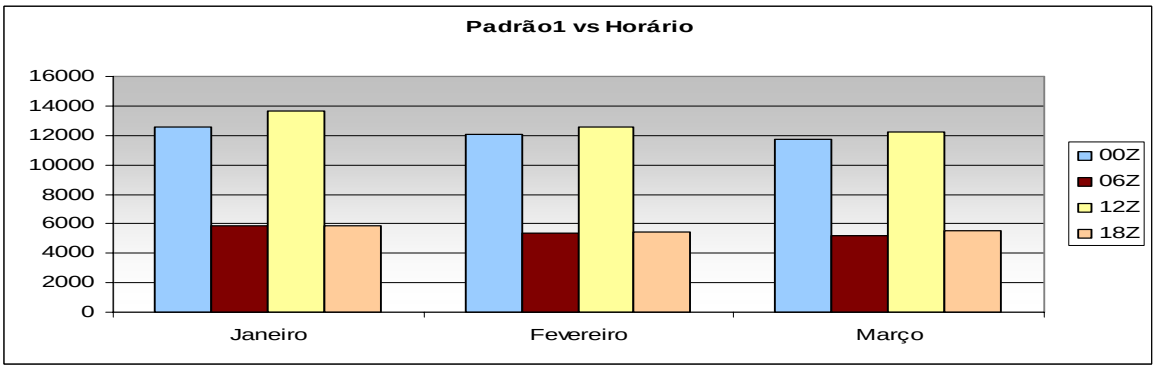

Figura 3. Avaliação dos erros em relação aos horários sinóticos

Verifica-se que a rejeição de dados ocorre em todos os horários sinóticos, sendo que a quantidade de erros é menor nos horários 06z e 18z. Esse fato deve-se a que a quantidade total de dados recebidos pelo sistema é também menor nesses horários.

Avaliação dos erros em termos do Instrumento Meteorológico

Os instrumentos meteorológicos avaliados são: plataformas terrestres, bóias marítimas, balões atmosféricos e outros. A figura 4 mostra a proporção de dados rejeitados em relação aos dados aceitos para um instrumento de Radiosondagem (Rawinsonde) e uma Plataforma Meteorológica embarcada em um navio (Surface Ship).

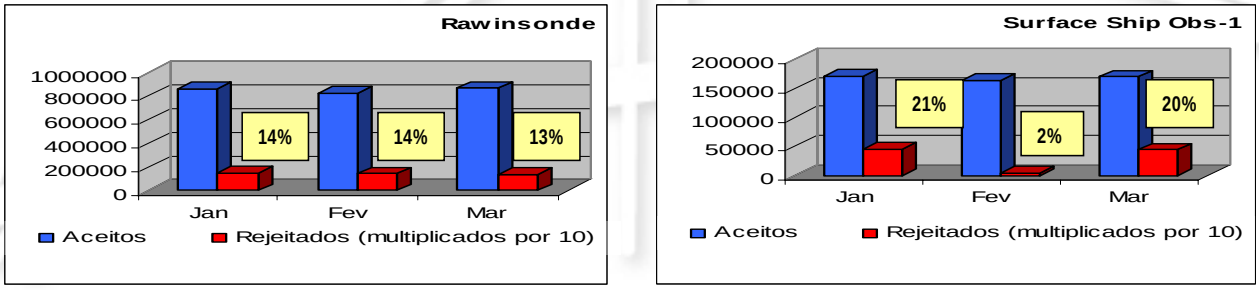

Figura 4 Quantidade de erros em relação ao Instrumento Meteorológico

Avaliação dos erros em termos da Variável Atmosférica

As variáveis meteorológicas analisadas em esta dimensão são: os ventos zonal e meridional, a altura geopotencial e o vapor de água. Os resultados obtidos para os ventos zonal no período de janeiro à junho são apresentados na figura 5.

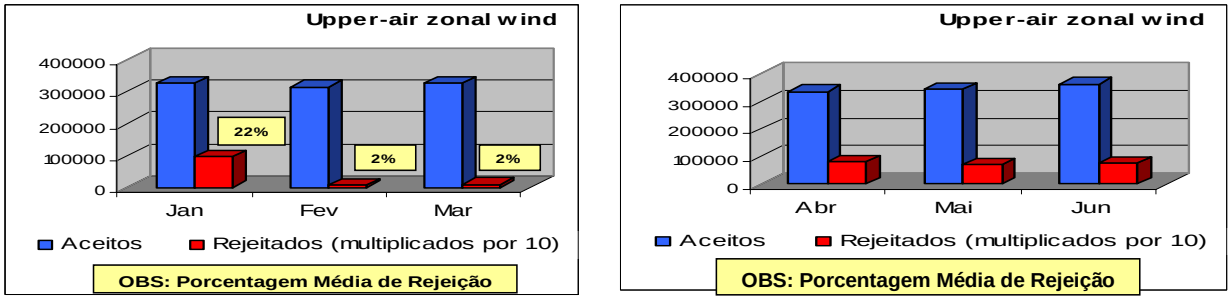

Figura 5 Quantidade de erros em relação à Variável Meteorológica

Avaliação dos erros em termos da localidade

Para realizar esta análise foi necessário segmentar o globo terrestre em várias regiões de interesse, como apresentados na figura 6.

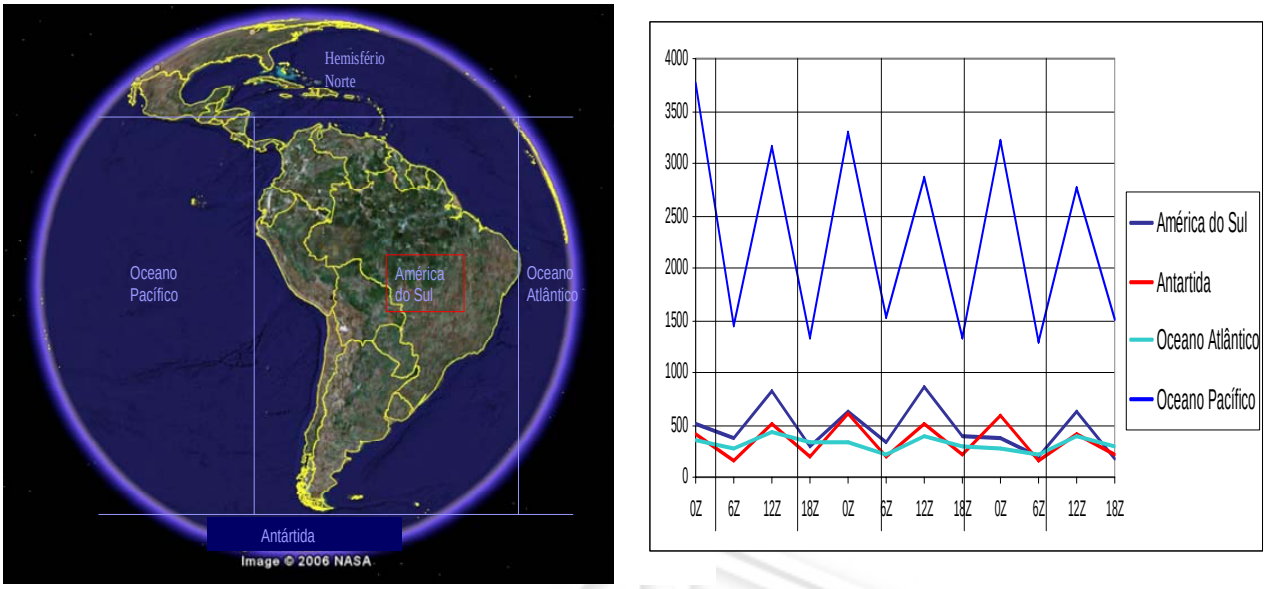

Figura 6. Regiões da Dimensão Localidade e Gráfico da Quantidade erros vs Localidade

O gráfico da figura 6 mostra a relação entre a quantidade de erros e as regiões: América do Sul, Oceano Pacífico, Oceano Atlântico e Antártida. O gráfico apresenta uma variação alternada da quantidade de erros, fato que confirma a menor quantidade de dados recebidos nos horários 06z e 18z como já foi discutido na seção 3.

4 Conclusões e trabalhos futuros

Apresentam-se aqui as principais conclusões em relação aos resultados obtidos pela análise das dimensões: período de tempo, instrumento meteorológico, localidade e variável atmosférica.

Na dimensão temporal verificou-se que os dados rejeitados ocorrem em todos os horários analisados e mantém uma ocorrência média independente do período de tempo estipulado. Em relação ao instrumento meteorológico, a análise mostrou que a média de erros observada, nem sempre se mantém no transcurso dos meses (veja a figura 4) e que o resultado depende do tipo de instrumento sob análise.

A análise na dimensão localidade constatou que a quantidade média de erros detectada varia de acordo com o tamanho e características de cada região, o que de alguma forma tem a ver com a quantidade de instrumentos meteorológicos existentes em cada uma delas. Assim, as áreas que apresentam menos erros são as áreas que incluem os oceanos e a região da Antártida. Em relação à

variável meteorológica, verificou-se que a quantidade de dados rejeitados apresenta algumas variações dependendo do período de tempo, o que pode significar a existência de uma componente sazonal neste tipo de avaliação.

Os resultados apresentados são preliminares, no entanto, percebe-se o potencial da análise dimensional para avaliar o comportamento da detecção de erros na Assimilação de Dados. Por exemplo, a ocorrência de erros persistentes no tempo, devido a anormalidades em instrumentos meteorológicos, certamente, será detectada pelo Data Warehouse , que poderá diagnosticar o tipo de instrumento e a localidade onde o mesmo se encontra.

O Data Warehouse permitirá realizar futuramente, no seu estágio operacional, a tarefa de estabelecer padrões de normalidade para a detecção de erros do sistema de Assimilação e realizar um controle periódico (diário, semanal, mensal) desses padrões de normalidade.

Um Sistema do tipo OLAP (On Line Analytical Processing) deverá ser incorporado ao Data Warehouse, para uma maior flexibilidade e dinâmica nas consultas que poderão incluir várias dimensões simultaneamente e diferentes níveis de detalhes na avaliação dos dados.

REFERÊNCIAS BIBLIOGRÁFICAS

Kimball, R. The Data Warehouse Toolkit. Guia completo para modelagem dimensional. Editora Campus, 2002.

Singh, S. Data Warehouse – Conceitos, Tecnologias, Implementação e Gerenciamento. Editora Makron Books. 2001.

Barbieri, C. BI – Business Intelligence. Modelagem & tecnologia. Editora Axcel Books, 2001.