VERIFICAÇÃO DA QUALIDADE DE DADOS ATRAVÉS DA LEI DE BENFORD

Fábio Favaretto (PUCPR)
fabio.favaretto@pucpr.br

Uma empresa de manufatura gera grandes quantidades de dados diariamente, sendo que estes serão processados e irão suportar decisões. Caso a qualidade destes dados não seja boa, as decisões podem ficar comprometidas. Dado um grande volume dee dados, é praticamente impossível a sua verificação detalhada e individual. Assim, existe a necessidade que seja feita uma verificação inicial do conjunto de dados que detecte inconsistências moderadas ou grandes. Caso isso seja percebido, podem ser feitas ações de descarte ou correção destes dados para impedir decisões também inconsistentes. O objetivo deste trabalho é testar a utilização de uma lei sobre a distribuição de dígitos significativos (Lei de Benford) para esta verificação inicial. Para este teste serão utilizados dados reais sobre a geração de refugos em uma empresa. Os resultados mostraram que pequenas alterações nos dados originais podem ser detectadas através desta utilização.

Palavras-chaves: Qualidade de dados, processo de decisão

1. Introdução

O mundo atual é caracterizado pelo uso intensivo de dados e informações, que são consumidos e gerados em grandes volumes. Uma empresa de manufatura gera grandes quantidades diárias de dados, como: todos os materiais recebidos, produzidos, entregues e movimentados dos estoques, entre vários outros. Estes dados são utilizados posteriormente para gerar relatórios e outras informações que irão servir de apoio a algumas decisões.

As empresas sofrem pressões para que seus processos de decisão sejam cada vez mais ágeis e flexíveis. Toda decisão requer um conjunto de informações. Estas devem ter a melhor qualidade possível, considerando aspectos (ou dimensões) como a acuracidade, precisão, consistência, acessibilidade e momento em que são disponibilizadas, entre outros. Para se aferir a qualidade de grandes volumes de dados ou informações são necessárias ferramentas e técnicas que fazem análises individuais e que, portanto demoram um tempo proporcional ao volume analisado.

A problemática que motivou esta pesquisa é a verificação inicial da consistência de dados que serão utilizados para suporte à decisão. Dado um grande volume de dados, é difícil a sua verificação detalhada e individual. Assim, existe a necessidade que seja feita uma verificação inicial do conjunto de dados que detecte inconsistências moderadas ou grandes. Caso isso seja percebido, podem ser feitas ações de descarte ou correção destes dados para impedir decisões também inconsistentes.

O objetivo deste trabalho é testar o uso da Lei de Benford como forma de fazer uma avaliação inicial de grandes volumes de dados. Esta lei foi proposta por F. Benford em 1938, como forma de reconhecer números anômalos. Segundo Hill (1996), a Lei de Benford trata da distribuição esperada para o primeiro e segundo dígitos significantes de uma seqüência de números (resultantes de medidas). Esta lei diz, por exemplo, que a quantidade de números que comecem com o dígito “1” é maior que aqueles que começam com o dígito “2” (ver Tabela 1) e assim sucessivamente. Como esta distribuição é conhecida, distribuições significativamente diferentes podem indicar algum problema, como por exemplo, a manipulação dos números.

A estrutura deste trabalho é apresentada a seguir. Após esta introdução é apresentada a metodologia de pesquisa utilizada, seguida pelos conceitos de qualidade de dados utilizados. Na Seção 4 são feitas considerações sobre o tratamento de dados com baixa qualidade e apresentada a Lei de Benford. Na Seção 5 são apresentados os dados utilizados para esta pesquisa, e na seção seguinte é apresentado o ambiente de testes utilizado. A Seção 7 apresenta os resultados obtidos e a Seção 8 finaliza o trabalho com as conclusões obtidas.

2. Metodologia

Quanto ao objetivo, a metodologia empregada nesta pesquisa é descritiva, que segundo Cervo e Bervian (2002) registra, analisa e correlaciona fatos ou fenômenos (variáveis) sem manipulá-los. Ainda segundo os autores, trata-se do estudo e da descrição das características, propriedades ou relações existentes na realidade pesquisada. O que se objetiva com este trabalho é estudar se a Lei de Benford tem a propriedade requerida para identificar dados inconsistentes em avaliações iniciais. Procura-se um resultado não absolutamente preciso, mas uma indicação de quando existe uma inconsistência moderada ou grande.
Para este desenvolvimento, a pesquisa foi desenvolvida nas seguintes etapas:

- Revisão bibliográfica sobre conceitos e a necessidade da qualidade de dados;

- Levantamento de pesquisas anteriores para identificação e tratamento de dados inconsistentes, entre eles a Lei de Benford;
- Aplicação da referida lei em dados reais e
- Análise dos resultados obtidos.

3. Qualidade de dados

Para o desenvolvimento deste trabalho, os termos "qualidade de dados" e "qualidade da informação" serão equivalentes.

Pesquisas realizadas na década de 90 do século passado revelaram que nos Estados Unidos, mais de 60% das firmas de médio porte com vendas anuais de mais de 20 milhões de dólares tinham problemas com qualidade da informação. Nos arquivos de registros criminais nos EUA, de 50% a 80% das informações são incompletas, ambíguas, ou sem acuracidade (WAND e WANG, 1996).

A base para o estudo da qualidade de dados e da informação é o conceito *de produto de informação* (PI), uma analogia direta com produtos resultantes de processos de manufatura tradicionais. Estes produtos "físicos" são resultantes de um processo de produção onde matérias primas são processadas. O produto de informação é resultante do processamento de dados (brutos) por sistemas de informação.

Arndt e Langbein (2002) entendem qualidade de dados como "encontro consistente das expectativas dos consumidores de informação".

As informações e os dados de uma organização são tratados como recursos, e por isso devem ser passíveis de mensuração, visto que os dados que são coletados, processados, acumulados e comunicados às empresas precisam ser mensurados de alguma forma (BEUREN, 2000). Guimarães e Évora (2004) afirmam que a informação é um recurso primordial para a tomada de decisão (no processo de produção). É inerente que esta informação tenha qualidade, no sentido de atender as necessidades do usuário.

Para Shankar e Watts (2003), a avaliação de critérios de QI é uma tarefa difícil. A avaliação deveria ser tão precisa e prática quanto possível. Este é um conflito de objetivos e um compromisso difícil de alcançar. Uma avaliação imprecisa pode tanto resultar em uma informação recuperada de baixa qualidade ou levar a evitar informações de alta qualidade. Avaliação imparcial também pode resultar em uma avaliação imprecisa ou levar a uma avaliação indevida de tempo e custo.

Conforme Wang e Strong (1996), três aproximações são utilizadas na literatura para estudar a qualidade de dados: uma intuitiva, uma teórica, e uma aproximação empírica. A aproximação intuitiva é feita quando a seleção da qualidade dos atributos dos dados para qualquer estudo particular é baseada na experiência dos pesquisadores ou por entendimento intuitivo a respeito de quais atributos são importantes. A maior parte dos estudos de qualidade de dados cai nesta categoria. Na literatura de sistemas de informação (SI), a qualidade da informação e a satisfação do usuário são as duas maiores dimensões para a avaliação do sucesso de um SI. Estas duas dimensões geralmente incluem alguns atributos da qualidade de dados, como a credibilidade, capacidade de tempo, precisão, confiabilidade, ocorrência, integridade (no sentido de estar completo), e relevância. Outros atributos como a acessibilidade e a interpretabilidade são também utilizados na literatura sobre qualidade de dados.

De acordo com Wang *et alli* (2000), muitas iniciativas corporativas, tais como o *Business-to-Business*, Gerenciamento Integrado de Cadeia de Suprimentos e o ERP (*Enterprise Resources Planning)*, correm o risco de falhar se não forem consideradas a qualidade e a melhoria dos dados. A incerteza a respeito da qualidade e as perdas financeiras decorrentes da baixa qualidade dos dados e informações, bem como a dificuldade em gerar transformações e análises fora do padrão, têm sido motivos de verdadeiros pesadelos organizacionais, ainda segundo Wang *et alli* (2000).

Vários estudos mostram que os dados armazenados na maioria dos bancos de dados das organizações não são sempre consistentes (acurados) e isto representa um grave problema, sendo que estimativas apontam estes erros entre 20 e 30% do total de dados armazenados (HASSAN, 2003).

4. Identificação e tratamento de dados com baixa qualidade

Quando discorre sobre erros em dados, Teng (2004) afirma que muitas vezes uma nova coleta dos dados errados é impossível. Então, antes de descartar estes dados, é necessário monitorar sua qualidade tanto quanto possível. O autor apresenta uma forma de filtrar dados que não podem ser recuperados, e um método de ajustar alguns dados, chamada de "polimento". Este método é baseado em algoritmos robustos, e possui aplicação limitada.

Favaretto e Mattioda (2005) propõen uma forma de medir a qualidade dos dados, o que permitiria identificar aqueles que não podem ser utilizados para compor informações mais agregadas e que não deveriam ser usados nas tomadas de decisão.

O trabalho de Benford sobre números anômalos é chamado de "Lei de Benford", e segundo Hassan (2003) trata da análise da distribuição dos primeiros e segundos dígitos de dados numéricos. Desvios desta distribuição indicam dúvidas quanto à consistência e autenticidade do conjunto de dados, requerendo uma investigação mais detalhada. Esta lei é válida para números que resultam do mesmo evento ou de fenômenos inter relacionados, como o comprimento de rios e populações de cidades. Nestes casos, é esperada uma freqüência maior do dígito "1" que qualquer outro dígito na primeira posição. A distribuição proposta por Benford para a distribuição dos primeiros e segundos dígitos é apresentada na Tabela 1. O dígito zero na primeira posição não é esperado por não ser significativo, porém na segunda posição possui uma distribuição maior que qualquer outro dígito.

Dígito	Primeira posição	Segunda posição
0	0,000	0,120
1	0,301	0,114
2	0,176	0,109
3	0,125	0,104
4	0,097	0,100
5	0,079	0,097
6	0,067	0,093
7	0,058	0,090
8	0,051	0,088
9	0,046	0,058

Tabela 1: Freqüências dos primeiros e segundos dígitos de acordo com a Lei de Benford (adaptado de HASSAN, 2003).

A utilização da Lei de Benford para identificação de dados inconsistentes pode ser automatizada, e por isso será utilizada para atender o objetivo deste trabalho. Hill (1996) faz um histórico da aplicação desta lei, apresentando também exemplos de sua aplicação.

5. Dados utilizados para teste da Lei de Benford

Os dados utilizados para verificação do objetivo proposto no trabalho são reais e foram obtidos diretamente no banco de dados do sistema ERP de uma empresa multinacional de autopeças. Estes dados são relativos ao controle diário de refugos de um dos produtos, especificamente a quantidade de produtos refugados e o custo dos mesmos. As quantidades foram obtidas por observação direta pelos próprios operadores que identificaram os problemas que levaram as peças a serem refugadas. O custo do refugo foi atribuído em função de uma série de fatores: custo das matérias primas (a maioria importadas), cotação de moedas estrangeiras, atribuições diretas e indiretas de custo e a operação que detectou o refugo.

O interesse em analisar estes dados vem do fato que parte da remuneração dos operadores é inversamente proporcional à quantidade de refugos. Os dados utilizados foram coletados entre Janeiro de 2001 e Outubro de 2003, sendo utilizados 34200 registros.

As quantidades de refugos observadas em produção normal são da ordem de grandeza de dezenas. Em alguns casos de teste de linha e ensaios destrutivos toda a produção de vários dias é refugada, sendo que nestas situações as quantidades podem chegar a dezenas de milhares. Vale destacar que entre os dados existia um valor cadastrado de 10 milhões de refugos para um único dia, e este registro foi descartado manualmente por ser impossível, visto que isto equivaleria à produção de dezenas de anos. Como vários registros apresentavam quantidades de refugos inferiores à dezena, não foi possível fazer a análise do segundo dígito.

Os custos dos refugos são registrados em Reais, e são necessariamente maiores que zero. Como alguns destes custos são inferiores à unidade, todos foram multiplicados por 100, o que não afeta os dígitos e posição; além disso, a rotina criada para separar o primeiro e o segundo dígito não correu o risco de separar uma vírgula no lugar de um dígito.

6. Descrição do ambiente de teste

Os dados utilizados foram importados para uma planilha eletrônica, onde foram realizadas todas as rotinas necessárias. Inicialmente foram feitas algumas preparações com estes dados, descritas na seção anterior.

Posteriormente, foi aplicada uma função da planilha eletrônica que retorna uma seqüência de caracteres de um dado qualquer, seja ele numérico ou não. A seguir foram criadas novas colunas com os dígitos analisados: apenas o primeiro para a quantidade de refugos e o primeiro e o segundo para o custo dos refugos. Cada uma destas novas colunas foi ordenada de forma crescente e computadas a quantidade de cada dígito em cada coluna. A ocorrência de zeros como primeiro dígito foi descartada, visto que isto significa que não ocorreram refugos para aquele produto naquele dia, e consequentemente também não houve custo. Assim, para cálculo das ocorrências (em porcentagem) dos primeiros dígitos foi considerado o total de dígitos de 1 a 9. Para o segundo dígito foi também considerado o dígito 0.

7. Análise dos resultados

Após a realização das atividades descritas na seção anterior, chegou-se aos resultados apresentados na Tabela 2.

	Quantidade de refugos	Custo dos refugos	
Dígito	**Primeiro dígito**	**Primeiro dígito**	**Segundo dígito**
0	0,000	0,000	0,116
1	0,436	0,290	0,105
2	0,202	0,171	0,105
3	0,123	0,166	0,126
4	0,078	0,074	0,092
5	0,051	0,078	0,099
6	0,040	0,081	0,120
7	0,030	0,061	0,085
8	0,024	0,047	0,076
9	0,016	0,033	0,077

Tabela 2: Freqüências da contagem de dígitos.

A Figura 1 apresenta os resultados obtidos para o primeiro dígito do custo de refugos e a comparação com a distribuição da Lei de Benford.

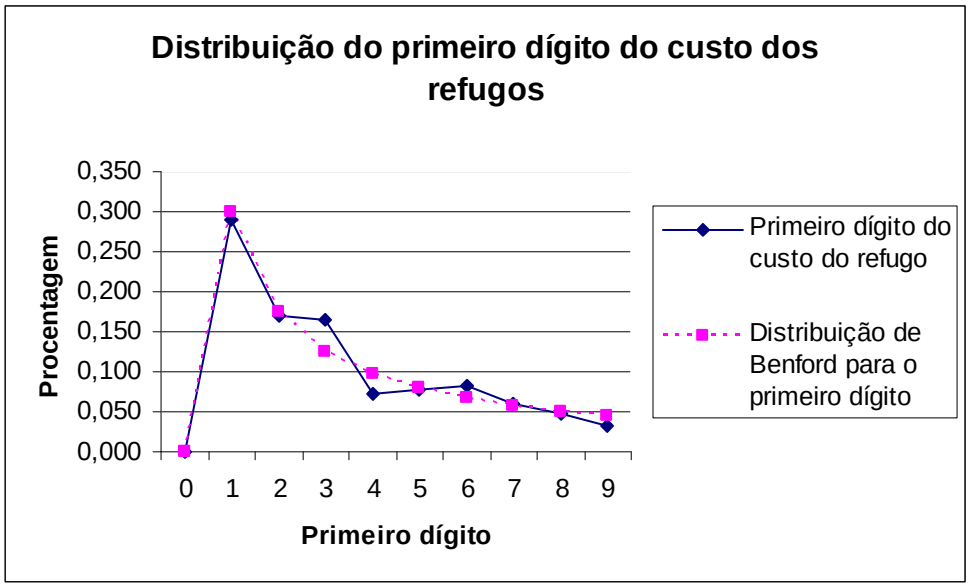

Figura 1 – Comparação da distribuição do primeiro dígito do custo de refugos com a distribuição esperada pela Lei de Benford.

A Figura 2 apresenta os resultados obtidos para o segundo dígito do custo de refugos e a comparação com a distribuição da Lei de Benford.

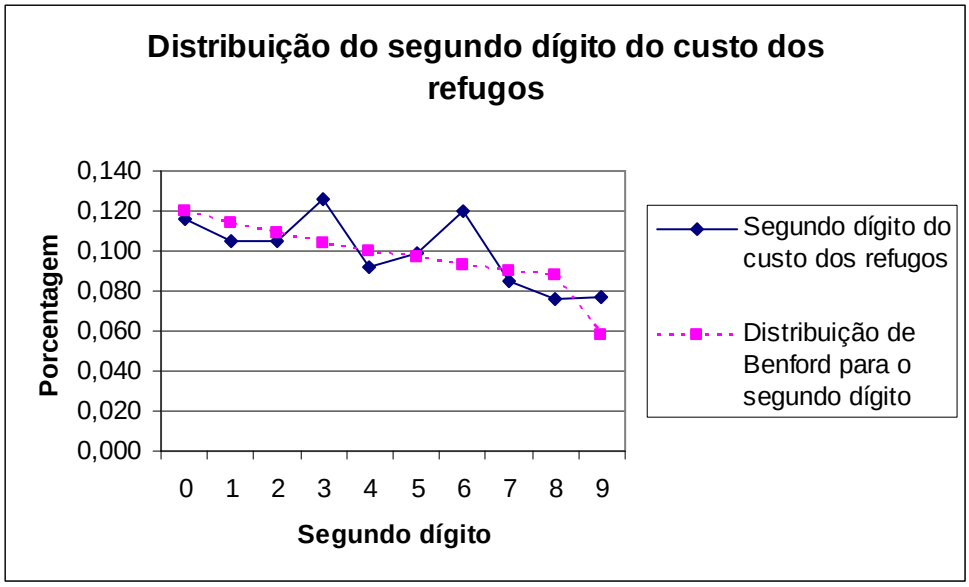

Figura 2 – Comparação da distribuição do segundo dígito do custo de refugos com a distribuição esperada pela Lei de Benford.

A Figura 3 apresenta os resultados obtidos para o primeiro dígito da quantidade de refugos e a comparação com a distribuição da Lei de Benford.

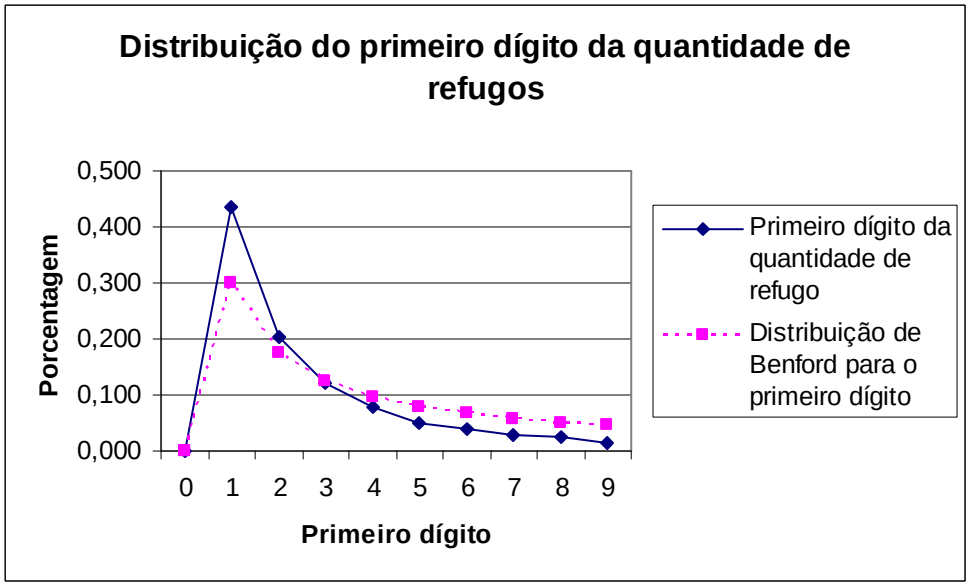

Figura 3 – Comparação da distribuição do primeiro dígito da quantidade de refugos com a distribuição esperada pela Lei de Benford.

Em uma análise inicial, as aproximações obtidas com as distribuições dos primeiros dígitos (Figuras 1 e 3) parecem razoáveis. Para se obter resultados mais significativos foram feitos testes de correlação entre as distribuições e calculados os desvios padrões.

Os coeficientes de correlação entre as distribuições obtidas e a esperada pela Lei de Benford são apresentados na Tabela 3.

	Quantidade de refugos	**Custo dos refugos**	
	Primeiro dígito	**Primeiro dígito**	**Segundo dígito**
Coeficiente de correlação com a distribuição esperada da Lei de Benford	0,9848	0,9792	0,6607

Tabela 3 - Coeficientes de correlação entre as distribuições obtidas e as esperadas pela Lei de Benford.

Os coeficientes obtidos para os primeiros dígitos indicam uma correlação muito forte. Isso significa que os resultados obtidos se aproximam bastante dos resultados esperados. A correlação para o segundo dígito do custo dos refugos apresenta uma correlação significativa, porém não tão forte como para os primeiros dígitos. Complementarmente foi feito cálculo do desvio padrão para as distribuições obtidas em comparação às esperadas, apresentado na Tabela 4.

	Quantidade de refugos	**Custo dos refugos**	
	Primeiro dígito	**Primeiro dígito**	**Segundo dígito**
Desvios padrão em relação à distribuição da Lei de Benford	0,1084	0,0831	0,0171

Tabela 4 – Desvios padrão entre as distribuições obtidas e as esperadas pela Lei de Benford.

Os desvios padrão apresentados mostram que as distribuições obtidas não estão distantes das distribuições esperadas.

Para atender o objetivo deste trabalho, foi desenvolvida uma análise de sensibilidade com os resultados obtidos. Com esta análise foi investigada qual deveria ser a variação nos dados originais para que uma análise inicial com base na Lei de Benford pudesse detectar problemas na qualidade dos dados.

Esta análise de sensibilidade ateve-se ao primeiro dígito do custo de refugos. O valor original obtido para a distribuição de dados que começam com o dígito “1” é de 0,290. Forçando uma variação de 0,001 (0,1%) para 0,291, o coeficiente de correlação passa de 0,9792 para 0,9795, e o desvio padrão passa de 0,0831 para 0,0832. Com isso, pode-se dizer que esta variação, que corresponderia à modificação de aproximadamente 34 dados, seria detectada. Isso equivale dizer que se 34 dos mais de 34000 registros fossem adulterados por algum motivo, isto poderia ser percebido.

8. Conclusões

Por se tratar de um estudo descritivo, as conclusões são limitadas e não podem ser generalizadas. Para o caso analisado, pode-se afirmar que a Lei de Benford mostrou-se uma boa aproximação para análises primárias da qualidade de dados, e que a adulteração de uma pequena quantidade de dados poderia ser detectada. Isso permite a construção de uma hipótese que poderia ser testada em investigações futuras, a de que a Lei de Benford é uma ferramenta útil para análises iniciais da qualidade de grandes volumes de dados. Para

verificação desta hipótese, entre outras providências, seria necessária uma definição dos limites aceitáveis para as variações entre as distribuições obtidas com aquelas esperadas, o que equivale à qualidade (atendimento das expectativas) desejada pelos usuários.

Referências

ARNDT D. & LANGBEIN N. Data quality in the context of customer segmentation. *Proceedings of the Seventh MIT International Conference on Information Quality (ICIQ-02)*, 2002.

BEUREN, I. M. *Gerenciamento da informação: um recurso no processo estratégico empresarial.* Editora Atlas, 2000.

CERVO, A. L. & BERVIAN, P. A. *Metodologia científica*. Quinta edição. Editora Prentice Hall, 2002.

FAVARETTO, F. & MATTIODA, R. A. A. Medição da qualidade da informação: um experimento na pesquisa em bases de dados científicas. *Anais do XXV Encontro Nacional de Engenharia de Produção – ENEGEP*, Porto Alegre, 2005.

GUIMARÃES, E. M. P. & ÉVORA, Y. D. M. Sistema de informação: instrumento para tomada de decisão no exercício da gerência. *Ciência da informação*, Volume 33, Número 1, Janeiro/Abril, 2004.

HASSAN, B. Examining data accuracy and authenticity with leading digit frequency analysis. *Industrial Management & Data Systems*. Volume 103, Número 2, 2003.

HILL, T. P. The first-digit phenomenom. *American Scientists*, No. 86, 1996.

SHANKAR, G. & WATTS S. A relevant, believable approach for data quality assessment. *Proceedings of the Eighth MIT International Conference on Information Quality (ICIQ-03)*, 2003.

TENG, C. M. Polishing blemishes: issues on data correction. *IEEE Inteligent Systems*, Vol. 4, 2004.

WAND, Y. & WANG, R. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, Vol. 39, Num. 11, 1996.

WANG, R. & STRONG, Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems*, Vol. 12, No. 4, 1996.

WANG, R., ZIAD, M. & LEE, Y. W. *Data Quality*. Kluwer Academic Publishers, 2000.