EDITORAÇÃO DE DADOS ERRÔNEOS EM SURVEYS

EDITING OF INCORRECT DATA EM QUANTITATIVE SURVEYS

EDICIÓN DE DATOS ERRÓNEOS EN ENCUESTAS

Henkel, Karl; Almeida, Jimnah de. (2023). Editoração de dados errôneos em surveys. Revista Gestão & Tecnologia. v. 23. nº 2. p. 81-104, 2023.

Karl Henkel
Professor Adjunto na Eberhard Karls Universität Tübingen
https://orcid.org/0000-0001-7032-2898

Jimnah de Almeida
Coordenadora do Departamento de Estatístico e Levantamentos Sistema Nacional de Emprego – SINE, Belém
https://orcid.org/0000-0003-2160-6557

Editor Científico: José Edson Lara
Organização Comitê Científico
Double Blind Review pelo SEER/OJS
Recebido em 22/12/2021
Aprovado em 30/05/2023

This work is licensed under a Creative Commons Attribution – Non-Commercial 3.0 Brazil

Resumo

Objetivo: O presente trabalho descreve as principais fontes de *data bias* ou dados errôneos que podem ocorrer na base de dados e identifica os seus efeitos na análise estatística.
Metodologia: Os dados foram levantados por meio da aplicação de um questionário com 800 entrevistadores para a criação de um banco de dados. Estes foram digitados por codificadores diferentes com vistas a diferenciar os dados errôneos. A análise é descritiva e analítica com aplicação de métodos quantitativos. Foram executadas pesquisas experimentais para identificar suas influências.
Originalidade: Os dados foram analisados relacionado à tipo de pergunta, escala, codificação e digitação, o que representa um aspecto não investigado ainda. A aplicação dos resultados possibilita reduzir falhas na base de dados na área de pesquisa, marketing ou estatísticas oficiais.
Principais resultados: Os resultados mostram que perguntas abertas, fechadas e escalas dicotômicas ou poliatômicas geram falhas de forma diferente. A identificação de dados errôneos por dupla digitação é restritiva pelo aspecto custo-benefício, a aplicação de lógicas algorítmicas é subjetiva e a substituição destes dados por outros pode criar a caraterização de dados manipulados.
Contribuições teóricas: Dados errôneos se previne intervalos mais longas ou uso de *tools* com sinais de vozes no momento da digitação. A aplicação de uma amostragem probabilística estratificada para a detecção de dados errôneos gera resultados satisfatórios em bases de *big data*.

Palavras-chave: dados errôneos; codificação de dados; inter-rater reliabilidade; editoração de dados.

Abstract:

Objective: This paper describes the main sources of data bias or erroneous data that occur in the database and identifies their effects on statistical analysis.
Methodology: The data were collected through the application of a questionnaire with 800 interviewers for creating a database. These were typed by different encoders in order to distinguish the erroneous data. The analysis is descriptive with application of quantitative methods. Experimental research was carried out to identify their influences.
Relevance: The data were analysed in relation to type of question, scale, coding and typing style, which is an aspect not investigated yet. The application of the results makes it possible to reduce failures in the database in the area of research, marketing or official statistics.

Main results: The results show that open-ended questions, closed questions and dichotomous or polyatomic scales generate failures differently. The identification of erroneous data by double typing is restrictive due to the cost-benefit aspect, the application of algorithmic logic is subjective and the substitution of these data by others can create the characterization of manipulated data.

Contributions: Erroneous data is prevented by longer pauses or use of tools with voice signals at the time of data entry. The application of a stratified random sample to detect erroneous data generates satisfactory results also in big data bases.

Key words: erroneous data; data codification; inter-rater reliability; data cleansing.

Resumen

Objetivo: Este trabajo describe las fuentes de sesgo de datos que pueden ocurrir en la base de datos e identifica sus efectos en el análisis estadístico.
Metodología: Los datos fueron recolectados mediante la aplicación de un cuestionario con 800 entrevistadores. Estos fueron mecanografiados por diferentes codificadores para diferenciar los. El análisis es descriptivo y con la aplicación de métodos cuantitativos. Se realizó investigaciones experimentales para identificar sus influencias.
Originalidad: La investigación analiza los datos erróneos en relación con la codificación, el tipo de pregunta, la escala y estilo de mecanografía, lo que representa un aspecto que aún no ha sido investigado. La aplicación de los resultados hace que sea posible reducir las fallas de la base de datos en el ámbito de la investigación, el marketing o las estadísticas oficiales.
Principales resultados: Los resultados muestran que las preguntas abiertas y las escalas poliatómicas generan más fallas. La identificación de datos erróneos por doble tipificación es restrictiva por el aspecto coste-beneficio, la aplicación de la lógica algorítmica es subjetiva y la sustitución de estos datos puede crear la caracterización de datos manipulados.
Contribuciones: Las pausas más largas reducen los errores de datos o utilizando herramientas con señales de voz en el momento de la entrada de datos. La aplicación de una muestra aleatoria estratificada para la detección de datos erróneos genera resultados satisfactorios también en grandes bases de datos.

Palabras clave: datos erróneos; codificación de datos; confiabilidad entre codificadores; limpieza de datos.

1. INTRODUÇÃO

A disseminação e a crescente facilidade no uso de sistemas computacionais trouxeram um avanço na pesquisa empírica, não somente porque possibilitaram tratar grandes quantidades de dados mensurados com ajuda do processamento computadorizado, mas também porque tornaram essa metodologia de pesquisa predominante. Ao comparar métodos qualitativos e quantitativos, muitas vezes esquece-se completamente de que na pesquisa quantitativa não são usados somente dados numéricos. Dados mensurados das Ciências Humanas e das Ciências Sociais Aplicadas diferenciam-se dos dados das Ciências Naturais, de maneira que não precisam ser somente mensurados, mas informados, porque se trata de ambientes comunicativos. O informante dos dados, o ser humano, tem uma vida subjetiva, uma

procedência e vivência social, o que influencia na compreensão dos fatos, na comunicação e na transferência dos dados. Embora hoje essa transferência seja feita de maneira informatizada, a aquisição dos dados mensurados em ambientes universitários, institucionais e em pesquisas quantitativas de pequeno porte representa ainda uma etapa manual. Neste processo, no ambiente computacional, há um número desconhecido de vieses ou bias de dados (*data bias*) (Lavalle, Maté, & Trujillo, 2020, p. 84) que não são originados de um incorreto levantamento, mas da aquisição dos dados e que dependem da habilidade do administrador, do tipo de pergunta, entre outros aspectos. Estas bias de digitação e transformação de dados geram erros que prejudicam a confiabilidade da pesquisa. No entanto, há pouco conhecimento sistemático desses fatores de influência e seus efeitos e precisam, portanto, de um realinhamento conceitual. Há necessidade de cuidados no processamento e na definição da qualidade dos dados, assim como no reconhecimento e na prevenção das bias de dados.

O objetivo deste trabalho é analisar e detectar as fontes e os processos que causam bias de dados ou dados errôneos, reduzi-los e identificá-los na base de dados, conhecer técnicas que permitam preveni-los e mostrar as consequências da sua permanência nos resultados. Ademais, são apresentados métodos de identificação viáveis e aplicáveis para pesquisas de pequeno e médio porte. Assim, aplica-se a análise em relação ao processo da digitação e ao tipo de pergunta e de escala, para obter um conhecimento funcional dos dados errôneos, o que faz parte do processo da validação dos dados (*reliability study*). Estes processos são analisados por meio dos dados levantados de um survey de porte médio[1]. Neste trabalho, compreende-se tratamento e processamento de dados como o momento a partir da captura visual ou verbal da criação de base de dados. Observando os trabalhos sobre este assunto, percebe-se que são poucos os que investigam a problemática de dados errôneos causados por este tratamento e processamento. Estes aspectos são de grande relevância para as instituições que levantam dados primários, aplicam métodos quantitativos em geral, preparam a informatização dos dados e os analisam estatisticamente.

Na introdução, são apresentados o conceito de dados errôneos, bem como a exemplificação do tratamento e da codificação com representações numéricas que podem

[1] Projeto IB20-010 – 2020, financiado pela Fundação Konrad Adenauer – KAS, Rio de Janeiro e Berlin.

ocasioná-los. Esses aspectos incluem identificar a fonte das causas dos dados errôneos na codificação e na digitação, analisando-os também na sua relação com o questionário, formas de perguntas e escalas. No final, são exemplificadas as consequências que os dados errôneos podem causar.

2. REFERENCIAL TEÓRICO

Dados levantados com intermediação comunicativa, ou seja, entrevistas, são obtidos, em geral, por meio de perguntas ordenadas em questionários, mensurados em escalas de intervalo, ordinal e nominal. Os dados são representados por valores numéricos e alfanuméricos, sendo estes últimos expressos com letras de forma discursiva. Os dados levantados representam ainda dados brutos, que precisam, ao contrário de dados secundários, ser organizados para futuro uso como base de dados em sistemas computacionais. A organização dos dados de forma lógica cria informações e conhecimento sistematizado.

Entretanto, qualquer tratamento e processamento de dados com envolvimento humano está sujeito a produzir falhas ou dados errôneos. A Ciência da Computação e Pesquisa em Banco de Dados definem dados errôneos como aqueles que ocorrem quando, no momento do processamento, não são aplicadas regras computacionais adequadas ou estão desatualizadas (Azeroual, Saake, & Abuosba, 2019, p. 4). Para a Ciência de Dados, dados errôneos indicam falhas no levantamento e são dados inconsistentes (Schwarz, 2018, p. 25), e representam, dessa forma, dados não seguros ou irracionais (Mao & Liu, 2014, p. 183) ou erros que correspondem a uma imprecisão de fatos reais com relação à representação esperada (Müller, Weis, Bleiholder, & Leser, 2005, p. 26). Sob a ótica da Estatística, dados errôneos representam erros algorítmicos (Waal, 2003, p. 11) e são declarados como dados suspeitos (*suspicious data*) (Granquist, 2011, p. 397). A Ciência Econômica entende dados errôneos como imprecisos, incompletos e inconsistentes (*erroneous data*) (March, 2005, p. 105). A Ciência Natural os define como aqueles valores que uma determinada variável não pode assumir (Silva, 2013, p. 22).

Embora Faulbaum (2014, p. 440), dentro da concepção de erro total de um survey, categorize dados errôneos como erros não amostrais, erros de observação e erros de processamento, as Ciências Sociais não apresentam definições específicas de dados errôneos e explicações sobre qualidade de dados (Yip, 2007, p. 368), porque aplicam mais métodos qualitativos na forma interpretativa e subjetiva do que pesquisas quantitativas na forma descritiva e objetiva com maior incorporação estatística.

Segundo a posição em que aparecem, os dados errôneos se classificam como: os originados do levantamento (*input error*), da aquisição (*acquisition error*), do momento do processamento (*processing error*) e os localizados dentro dos resultados (*output error*).

Em ambientes, em que ocorre o processo interativo entrevistador versus entrevistado (*face-to-face*), os dados são transmitidos por meio da comunicação oral ou visual e o valor pode ser capturado pelo entrevistador de forma errada, ou o entrevistado, no caso de autoadministração do preenchimento do questionário, aloca o dado na posição errada no questionário (*input error*). Entretanto, respostas falsas, socialmente desejadas ou dados comunicados que representam supostamente valores irracionais ou não refletem um fato real, quando são dados conscientemente e não se tratam como dados por engano, expressam para o entrevistador dados errôneos, mas uma decisão racional para o entrevistado em responder dessa forma. Assim, representam valores verdadeiros (Möhring & Schlütz, 2019, p. 42) e, portanto, não necessariamente dados errôneos.

Dados levantados por meio de questionários online, de dispositivos móveis, como *Touch-tone Data Entry* – TDE, de formas de preenchimento autoadministrado pelo entrevistado, como *Computer-Assisted Self-Administered Questionnaire* – CASAQ ou *Computer Assisted Personal Interviewing* – CAPI, representam respostas não necessariamente verdadeiras, mas dados corretos, quando alocados de modo consciente.

Da mesma forma, dados numéricos e alfanuméricos adquiridos de maneira digital por escaneamento com sistemas de reconhecimento ótico de caracteres (*Optical Character Recognition* e *Intelligent Character Recognition* − OCR/ICR) podem representar dados errôneos quando escaneados na posição incorreta ou de forma descaracterizada devido à impureza do escaneamento (*acquisition error*).

Pelo processamento computacional subsequente, dados podem ser operados de forma incorreta (*processing error*), causando erros diretamente na base de dados, além de ocasionarem falhas nos processos das análises estatísticas (*output error*). No entanto, dados errôneos podem criar um efeito compensatório entre si sem consequências negativas nos resultados.

A respeito da frequência e da reprodutibilidade de dados errôneos, Rahm and Do (2000, p. 3) diferenciam dados errôneos singulares dos que se acumulam. A última forma pode ser entendida como erro contínuo quando um erro singular reproduz, por meio de processamento, outros dados errôneos (*processing error*).

De acordo com a probabilidade, dados errôneos são definidos como aleatórios ou podem surgir em qualquer parte da base dos dados.

Em relação ao tempo para a identificação de dados errôneos, Liua, Andrienko, Wu, Cao, Jianga, Shi, Wang, and Hong (2018, p. 192) afirmam que a identificação e a correção de dados errôneos dentro da concepção de gestão da qualidade dos dados ocupam de 30 % a 80 % do tempo gasto nas tarefas. Manrique-Vallier and Reiter (2017, p. 1714) citam que em surveys 22 % dos dados errôneos ficam não identificados, causam as falhas estatísticas e concluem que uma contínua identificação de todos os dados teria um efeito muito reduzido nos resultados finais. Em projetos de pesquisas predomina a exatidão dos resultados com referência a uma ou algumas questões específicas, enquanto que no caso de estatísticas oficiais importa a atualidade dos dados de longo prazo, que pode ser prejudicada com um gasto excessivo de tempo de editoração para a correção dos dados errôneos (Brislinger and Moschner, 2019, p. 106), o que ocasiona atraso na entrega dos resultados e desatualização das informações nos dados secundários e oficiais.

A decisão de dar mais atenção para a identificação e correção dos dados errôneos mostra uma atitude científica mais criteriosa, porque, segundo levantamentos de Seligman, Rosenthal, Lehner and Smith (2002, p. 7), representa para pessoas com esta atitude o segundo processo mais intensivo em tempo durante o levantamento e tratamento de dados, enquanto pessoas menos criteriosas interpretam este processo como o mais difícil. A problemática faz parte da gestão da qualidade de dados (*data quality management;* DQM) (Waal, 2003, p. 11).

3. METODOLOGIA

Analisou-se a problemática de dados errôneos em um levantamento censo escolar. A base de dados é formada por 800 pessoas, entrevistadas em escolas públicas da rede estadual de ensino no município de Belém e selecionadas por meio de uma amostra aleatória simples. O levantamento dos dados ocorreu nas salas de aula mediante entrevistas com preenchimento do questionário, sob autoadministração do entrevistado. Os questionários foram analisados e os dados digitados manualmente.

Para se fazer uma comparação quanto à habilidade da digitação, foram selecionados quatro estudantes universitários da disciplina "Análise de Dados". Estes alunos receberam treinamento em metodologia técnico-operacional de tratamento de dados e orientação sobre o plano de codificação ou como computar dados em planilhas de banco de dados. No caso de perguntas abertas, a maior atenção foi dada à eliminação de uma interpretação subjetiva em relação às respostas e à criação de um conceito objetivado. Foram realizadas reuniões iniciais e, quando surgiram dúvidas por parte do aluno digitador, essas foram analisadas em conjunto. Depois, os alunos iniciaram, individualmente, os processos de codificação e de digitação dos dados na planilha computacional. A digitação dos dados foi analisada segundo as variáveis tempo de digitação e uso do teclado. Assim, os digitadores foram acompanhados por meio de um processo de treinamento e supervisão.

Sem a possibilidade de realização de reentrevistas e de correção das respostas, o questionário do survey foi o único documento que permitiu a reprodução das respostas e a geração das informações. Cada questionário contém 63 campos na planilha computacional do SPSS, 97 bits ou algoritmos e, devido à existência de códigos com mais de um algoritmo, o processo necessitou de até 100 teclas pressionadas para digitar os dados ou ca. 77.600 bits no total.

Embora seja comum que os próprios digitadores realizem a identificação e as correções dos dados errôneos, os dados dos questionários digitados em blocos ou lotes de 20 questionários por dia, com incorporação de intervalos de repouso, foram analisados pelo supervisor com o objetivo de padronizar a identificação e a correção dos dados errôneos, para que cada digitador não aplicasse uma sistemática identificadora própria.

Para a identificação dos dados errôneos, aplicou-se uma listagem da frequência dos códigos e valores para cada variável a fim de identificar, por meio da observação, os dados aquisitados fora dos limites (*wild codes*), levando em consideração variáveis mensuradas com escalas diferentes e respectivos códigos, segundo um plano de codificação, e um controle por meio de amostragem.

Para melhor entendimento do efeito que dados errôneos numa base de dados podem causar, aplicou-se um experimento com manipulação dos dados em uma variável dicotômica, a fim de identificar a relação entre dados errôneos e amostra. Outra experimentação foi aplicada para mostrar os efeitos da substituição de dados errôneos em surveys. Os dados foram digitados na planilha do programa SPSS.

4. APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

4.1 A concordância entre as codificações

O plano de codificação tem a função de orientar o codificador sobre a atribuição de certo código para a respectiva categoria de uma resposta dada em uma pergunta aberta. Estes dados precisam ainda de uma adequação para o futuro processamento pela categorização (Henkel, 2017). Neste processo, os diferentes códigos para categorizar respostas parecidas por um codificador não são tratados como dados errôneos, mas como variações da codificação ou como a habilidade do codificador em repetir as suas próprias decisões para alcançar alta reprodutibilidade das codificações (Krippendorff, 2004, p. 214), o que representa a confiabilidade intracodificador (*intracoder reliability*) (Graber, 2004, p. 55). Estas codificações, quando são realizadas por codificadores diferentes, são definidas como confiabilidade intercodificador (*intercoder reliability*) e também podem variar.

A concordância das codificações para dados mensurados em escalas numéricas é expressa com o coeficiente de Cronbach (r) e, no caso de escalas alfanuméricas, com o coeficiente Cohen's kappa (κ) (Cohen, 1960)[2]. O valor de κ depende do sistema categorial, seja

[2] Kappa é calculado da seguinte forma: $\kappa = (p_0 - p_e) / (1 - p_e)$, sendo p_0 o número de concordância entre os codificadores; e p_e a probabilidade hipotética estimada que a concordância pode alcançar em cada campo da matriz. Ver também González-Prieto, A., Perez, j., Diaz, J., & López-Fernández, D. (2020).

ele dicotômico, tricotômico ou múltiplo, do número das codificações relacionadas ao tamanho do survey e da potência estatística das categorias ou da forma como elas são distinguíveis umas das outras. Quanto maior o valor de κ, maior a distinguibilidade entre as categorias.

Por exemplo, a pergunta "*O que você acha sobre a democracia*?" pode ser categorizada de forma dicotômica, em que o código "*1*" representa respostas com um conteúdo "*positivo*" e o código "*2*", respostas com conteúdo "*negativo*". A digitação de um dado código não estabelecido pelo plano de codificação resulta num dado errôneo. A codificação "*4*", por exemplo, aplicada pelo codificador$_1$, mudaria uma matriz original de 2 × 2 com os códigos "*1*" e "*2*" (codificador$_1$ "*1*" / "*2*" e codificador$_2$ "*1*" / "*2*") para uma de 1 × 3 para o codificador que digitou o dado errôneo (enquanto o codificador$_2$ ficaria com a matriz 1 × 2 (codificador$_2$ "*1*" / "*2*"). Esta situação violenta o pré-requisito da disponibilidade e uso dos códigos por todos os codificadores e não se pode mais falar de pares de códigos iguais entre os codificadores. Dados errôneos têm consequências diferenciadas numa codificação (Tabela 1).

Tabela 1
Tamanho do survey, quota de concordância em %, Cohen's kappa κ sem e com dados errôneos

tamanho survey	quota de concordância	κ	nº de dados errôneos		
			1	2	3
			κ		
n = 50	80	0,541	0,505	0,469	0,419
n = 100		0,350	0,335	0,321	0,325
n = 200		0,487	0,476	0,464	0,456
n = 400		0,585	0,586	0,587	0,589

Fonte: Elaborado pelos autores.

A digitação de um dado errôneo num campo com concordância entre codificador$_1$ e codificador$_2$ ou discordância tem consequências opostas e κ pode tanto aumentar como diminuir.

4.2 Os dados errôneos em relação ao tempo de digitação, escalas e com o tipo de pergunta

Entre os digitadores não houve grandes diferenças no processo de digitação de dados errôneos, provavelmente em virtude do treinamento. A maioria dos dados digitados erroneamente foi identificada no 2º intervalo temporal da digitação ou quando esta ocorreu mais

rapidamente. Observou-se que, no início do processo, os digitadores fizeram uma identificação do posicionamento da pergunta no questionário (processo sensorial−visual) e da resposta (processo sensorial-visual), iniciaram a compreensão da resposta (processo cognitivo), identificaram a respectiva categoria (processo cognitivo), compartilharam o respectivo código com a categoria (integração sensorial−visual−cognitiva), localizaram o respectivo campo na planilha computacional (processo sensorial−visual), reconheceram a respectiva tecla (processo sensorial−visual) e finalmente digitaram o dado (processo táctil).

Após a fase inicial, os digitadores conseguiram agilizar o processo em razão das habilidades motoras adquiridas (figura 1). Quando a identificação visual das respostas e a habilidade táctil no âmbito do teclado se dão de forma mais fluida, pode ocorrer uma desintegração destas informações (Cunha, Bastos, Veiga, Cagy, Mcdowell, Furtado, Piedade, & Ribeiro, 2004, p. 667). A respeito da posição no questionário, 20 % dos dados digitados errados foram encontrados no 1º quartil, 35 % no 2º, 25 % no 3º e os 20 % restantes no 4º quartil das respostas do questionário (Figura 1).

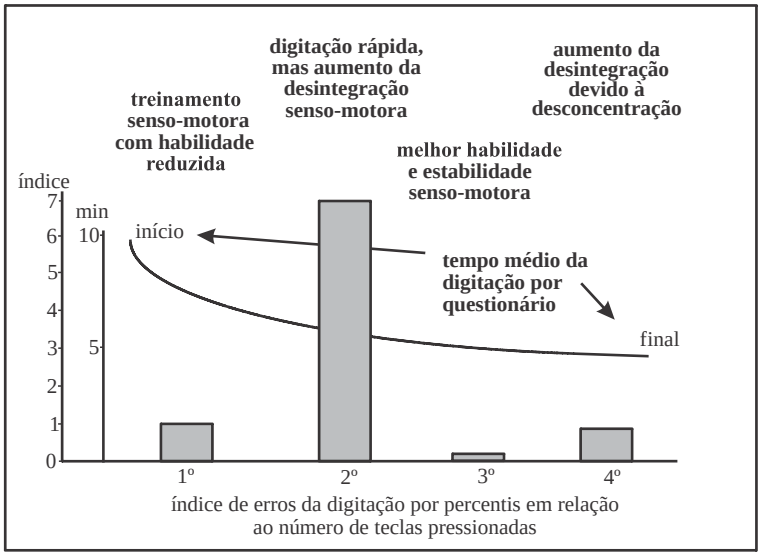

Figura 1: Índice de dados errôneos por quadris e tempo de digitação em minutos
Fonte: elaborada pelos autores.

Supõe-se que a baixa quota de dados errôneos ocorreu também pela permanente mudança do posicionamento de perguntas abertas e fechadas e pelo uso de escalas variadas no questionário, o que causou uma contínua reorganização cognitiva e captura sensorial-visual dos dados na fonte original pelo digitador.

Uma das fontes principais de dados errôneos foi a digitação de dados numéricos de blocos sequenciais de perguntas com escalas de intervalo. Uma digitação inicialmente lenta mudou para uma elevada habilidade motora e maior integração sensorial-motora, o que causou uma digitação sem aguardar o *feedback* táctil ou reconhecimento de que a tecla foi pressionada, ocorrendo uma desintegração dos dados seguintes. Esse processo pode causar uma digitação de dados de forma correta, mas na posição errada, o que cria dados errôneos em blocos inteiros e representa erros espaciais (Lindsey & Logan (2019, p. 397).

Segundo o tipo de escala, dados mensurados com escalas nominais causam mais dados digitados erroneamente (Figura 2).

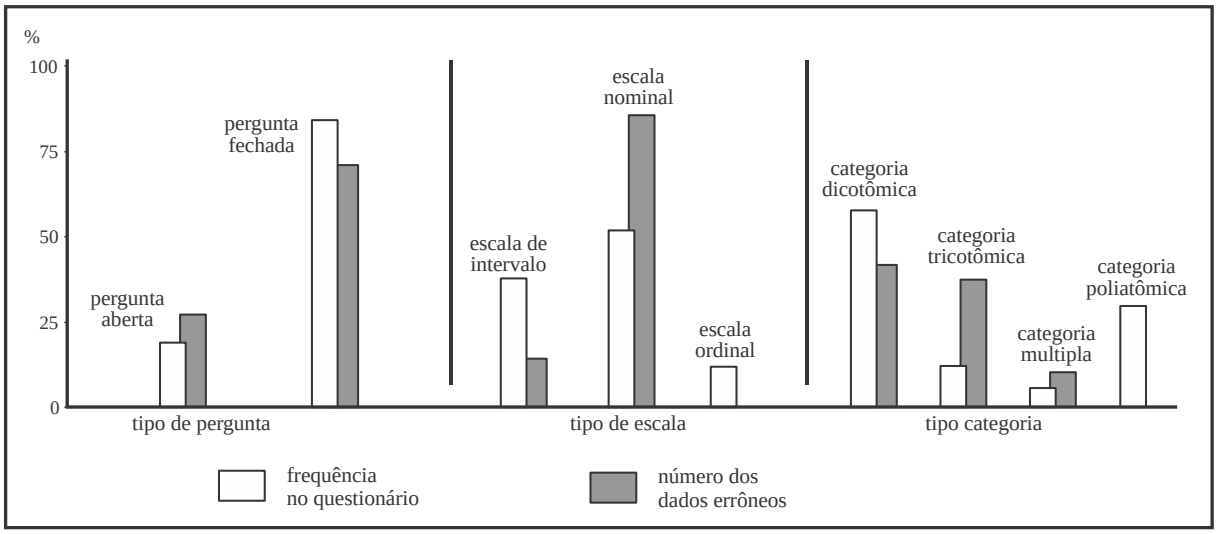

Figura 2: Tipo de pergunta, escala, categoria e índice de dados errôneos
Fonte: elaborada pelos autores.

Especificamente escalas nominais ≥ a três categorias causam mais dados errôneos, ao contrário das escalas ordinais, que têm em geral menos atributos categóricos. Nestes casos, a digitação de dados errôneos ocorre com maior frequência apenas quando o digitador inicia o

processo cognitivo-memorial na identificação do respectivo código dado, sem acompanhar, pelo processo sensorial−visual, a identificação da resposta no plano de codificação, ou quando não compartilha a identificação sensorial−visual da posição do atributo no questionário com o *feedback* físico−táctil. A Psicologia explica que isso ocorre porque há uma desativação parcial da memória em trabalhos automatizados como dados apresentados sequencialmente ("*1*" / "*1*" / "*1*" / "*1*" / "*2*"...) (Lau, 2018, p. 115) ou porque o digitador confunde visualmente a posição da resposta e digita um código-dado errôneo.

A respeito do tipo de perguntas, percebe-se que a pergunta aberta é mais sujeita à digitação de dados errôneos. A pergunta fechada, caracterizada pela disponibilização dos atributos das respostas, somente causa problemas para o digitador quando tem mais de dois atributos disponibilizados na escala nominal. Nestes casos, repete-se a condição de que quanto maior o número de dados que o digitador precisa cognitivamente processar, menores a memorização e a coordenação com o processo motor.

4.3 Métodos para a prevenção e identificação de dados errôneos

Uma prática para prevenir e identificar dados errôneos é a dupla digitação, que consiste na digitação dos dados por pessoas diferentes e posterior comparação, mas não é uma maneira eficiente para grandes amostras, embora eficaz na identificação de dados errôneos. Uma digitação de 800 questionários, por exemplo, considerando uma jornada de trabalho de quatro horas, ou vinte questionários por dia, tempo recomendado por Scheiner, Sicks and Holz-Rau (2014, p. 4) para níveis universitários, necessitaria de uma digitação em 40 dias. No nível comercial, com uma jornada de trabalho de oito horas por dia e com dois digitadores, o processamento dos dados e a posterior identificação e comparação demorariam vinte dias. Ambos os casos, porém, são irracionais.

Outro método é a utilização de programas ou *tools* módulos específicos para a digitação da entrada dos dados, que visualizam a digitação de um dado fora dos limites estabelecidos. Porém, uma digitação de dados errôneos pode ocorrer dentro dos limites.

Em relação ao teclado e à digitação, há a possibilidade de escolher posições de teclas que previnem a digitação de dados errôneos, porém com efeito de digitação mais lento, e outras que

aumentam a rapidez da digitação, todavia causando o aumento da digitação de dados errôneos. Com o objetivo de preveni-los, podem-se usar códigos que assegurem teclas com disposição espacial bem separada entre si (Medeiros, 2005, p. 11), como, por exemplo, "*1*" para "*sim*", "*3*" para "*não*" e "*9*" para "*não respondeu*", no caso de variáveis com categorias dicotômicas. Para agilizar o processo de digitação, podem-se usar teclas próximas umas das outras, sendo "*1*" para "*sim*", "*2*" para "*não*" e "*0*" para "*não respondeu*". Porém, a posição das teclas dessa maneira aumenta o risco de se digitar erroneamente, uma vez que a proximidade destas contribui para facilitar um engano sensorial−visual. Especificamente a digitação com alto nível de rapidez faz com que a tecla próxima da tecla pretendida seja pressionada acidentalmente, digitando-se o dado erroneamente.

Observou-se que, quando foi interrompida esta lógica por uma de "*0*" para "*não tem filhos*", "*1*" para "*um filho*", "*2*" para "*dois filhos*", etc. e adicionalmente "*9*" para "*sem resposta*" (*missing value code*), ocorreu um aumento da digitação de dados errôneos, provavelmente devido ao aumento das informações que precisam ser cognitivamente tratadas.

A procura sensorial-visual do código no plano de codificação para a respectiva resposta no questionário – cujo valor pode ser memorizado cognitivamente − e a procura sensorial-visual da respectiva tecla para digitar este código são, apesar do treinamento, aspectos de lentidão, e a habilidade movedora dos dedos entre as respectivas teclas caracteriza rapidez no processo da digitação.

Alguns alunos digitadores usaram notebooks com teclados soft (*rubberdome keyboard*); outros utilizaram computadores com teclas mecânicas. Os alunos do primeiro grupo digitaram mais rápido, provavelmente pela maior proximidade entre tela de computador e tecla, o que proporcionou uma melhor integração sensorial−visual e, como consequência, maior habilidade motora, no entanto, como efeito, digitaram mais dados errôneos.

Estes efeitos de *feedback* visual (*visual feedback*), *feedback* táctil (*tactile feedback*) ou sistemas modernos usam hoje sinais de vozes (*speech signal*) que emitem sinais acusticamente quando a tecla pretendida é pressionada (Rabin & Gordon, 2004, p. 367) e reduzem a digitação de dados errôneos.

Waal (2003, p. 11) apresenta alguns métodos para identificar dados errôneos em surveys, como: editoração computacional auxiliada por computadores (*computer-assisted editing*),

análise de resultados incomuns com impacto profundo nos resultados (*selective editing*) e análise da apresentação gráfica do resultado.

Programas computacionais conseguem, por meio de regras de operações lógicas (*matching rules*), identificar dados errôneos satisfatoriamente. Porém, a maneira de estabelecer as regras computacionais procedimentais e a decisão da continuação ou interrupção deste processo da editoração automática são individualmente determinadas, ou seja, subjetivas. Ademais, a relação input ou tempo investido, medido por meio do tempo aplicado na criação da regra computacional e posterior análise dos dados, e output ou resultado esperado, mensurado pelo número de dados errôneos identificados, é vantajosa somente para grande base de dados. Além disso, a editoração automática de dados alfanuméricos necessita, como pré-requisito, da codificação destes, que não gera dados errôneos, mas variações de dados dentro do plano codificador. Este método não consegue identificar um dado errôneo dentro dos limites das especificações. Na codificação, por exemplo, a categoria "*masculino*" recebeu o código "*1*", "*feminino*" "*2*" e "*não respondeu*" "*0*" e, dessa forma, qualquer outro código representou um dado errôneo. Porém, esta listagem não identificou uma digitação inversa, ou seja, "*2*", embora devesse ser codificado com "*1*", e os erros persistiram na base de dados.

O método da aplicação de testes de razoabilidade (*reasonableness tests*) verifica se os valores levantados estão congruentes com valores que podem ser esperados por meio de consulta de fontes de dados secundários de outras pesquisas (Winkler & Chen, 2001, p. 2). Entretanto, não se determina quais fontes devem ser consultadas e o que são valores esperados.

Outro método é o estabelecimento de faixas de valores ou quantis, nas quais os dados podem ser agrupados (Hwang, Kim & Jung, 2018, p. 243). Porém, a definição dos limites superiores e inferiores destas faixas de valores é determinística e varia também conforme a subjetividade.

O método gráfico ou identificação de valores extremos discrepantes (*outliers*; *inliers*) (Broeck, Cunningham, Eeckels, & Herbst, 2016) é aplicável para dados numéricos, mas não para dados alfanuméricos, e não faz distinção entre valores atípicos, que podem ser verdadeiros ou errôneos.

Maletic and Marcus (2000, p. 4) apresentam ainda como método a comparação do posicionamento dos dados em relação às medidas estatísticas média, desvio padrão,

abrangência e análises múltiplas, como a análise de cluster, etc. Hara, Nitanda and Maehara (2019, p. 2) aplicam uma análise de teste t para definir o aspecto de diferenças significantes entre dados corrigidos e não corrigidos. porém, este método não considera a editoração de dados alfanuméricos, que não são representáveis por medidas estatísticas.

Além desses métodos, existe o da validação externa ou comparação de dados levantados com dados das estatísticas oficiais ou dados secundários. Entretanto, surveys levantam, em geral, dados primários, indicando que fontes oficiais não dispõem destes dados ainda, o que exclui o método da comparação. Dados oficiais também podem ter falhas e são classificados por Braga, Lima, Leiva e Nascimento (2008, p. 7) como de alta, média e baixa confiabilidade.

A escolha de um método adequado para a presente pesquisa se orientou nas recomendações de Liao, You, and Zhang (2019, p. 379), que recomendam uma identificação dos dados de forma rápida e eficiente e com o mínimo de perda das informações. Para identificar os dados errôneos por meio desta sistemática, aplicou-se o método da otimização amostral (*mathematical optimisation*) e se estabeleceu um número máximo de erros que se pretende identificar. Foi determinada uma quota de identificação máxima de 98 % ou erro amostral de 2 % e um grau de significância de 99 % (z = 2,58), o que resultou numa necessária verificação em 67 campos na planilha computacional do banco de dados por questionário, ou seja, um total de 53.600 campos para 800 questionários e respectiva comparação com a fonte original, para alcançar esta medida estatística operacional.

Independentemente da determinação da quota máxima de dados errôneos que se pretendeu identificar, aplicou-se um controle sistemático. Foi aleatoriamente escolhido cada trigésimo segundo questionário, cujos valores anotados foram comparados com os valores digitados na planilha da base de dados. Posteriormente, fez-se uma comparação das variáveis que foram escolhidas aleatoriamente, ou seja, o dado de cada vigésima terceira variável de cada questionário foi comparado com o respectivo dado digitado na planilha. Depois, foram analisados dados mais sujeitos a serem digitados erroneamente, como dados levantados com escalas numéricas sequenciais. Novamente, por meio de uma escolha aleatória sistemática, foram comparados os dados de cada oitava pergunta que tinha uma escala numérica com os respectivos quatro dados seguintes na planilha para identificar erros sequenciais. Finalmente, compararam-se os dados de cada nona resposta do questionário com os respectivos códigos

desta resposta representados no plano de codificação. Este procedimento ajudou a detectar erros sistemáticos e transversais (Henkel & Almeida, 2003, p. 35). Por meio desta sistemática, foram identificados 157 dados errôneos, o que representa uma quota de 0,2 %, considerada desprezível quando se compara com a ideia de Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 28), que estimam um erro em até 10 % dos dados em surveys como rotineiros. Já Schomburg (2001, p. 66) considera um índice de erro de 2-3 % como aceitável. Para Fowler (2008, p. 152), bons digitadores alcançam menos de um erro digitado entre 1.000 bits (= 0,1 %), o que representaria 76 dados.

Entretanto, há dados errôneos em condição de erro latente (Tröger, 2018, p. 84) que são falhas impossíveis de serem identificados sem um controle total da base de dados e fonte original (Boeschoten, Oberski, & Waal, 2017, p. 922), uma vez que estes dados podem representar um valor dentro das faixas estabelecidas, ou uma comparação dos resultados com dados oficiais pode mostrar que estão dentro das expectativas.

4.4 O impacto de dados errôneos nos resultados e os métodos de correção

A validação é o processo de verificação da usabilidade da base de dados e de reconhecimento de dados errôneos (Zio, Fursova, Gelsema, Gießing, Guarnera, Petrauskienė, Kalben, Scanu, Bosch, Loo, & Walsdorfer, 2016, p. 61). Isso implica também entender a função destes, por exemplo, nas consequências que podem causar nos resultados finais (Saha & Srivastava, 2014, p. 1294). Dentro do conceito da survey experimentação, Saam (2015, p. 4) usa o termo simulação, o que é interpreta como a relação de um modelo com os dados e, portanto, com o computador. Para melhor interpretar os impactos de dados errôneos nos resultados, foram manipulados dados da variável dicotômica "*sexo do entrevistado*". Desta maneira, de n = 782 declarações válidas (18 dados = *missing value*) foram digitados n = 24 dados errôneos de maneira artificial (= 3,1 % da base de dados desta variável), de modo que os códigos-dados foram digitados de forma inversa, sendo que, ao contrário dos dados verdadeiros de "*2*" para "*feminino*", foram digitados com o código "*1*" como se fosse "*masculino*". A experimentação mostrou que, no caso de um survey com um erro amostral de, por exemplo, ± 2,0 %, considerado comum nas Ciências Humanas, a partir de 18 dos 782 dados ou 2,3 % dos

dados erroneamente aquisitados, ocorre uma inversão do resultado. Nesse caso, uma expressão como "*A maioria dos entrevistados é do sexo feminino*" mudaria para "*A maioria dos entrevistados é do sexo masculino*" (*goodness of distributional fit*) (Rohwer, 2014, p. 320), o que leva o processo da tomada de decisões a outras determinações (Figura 3).

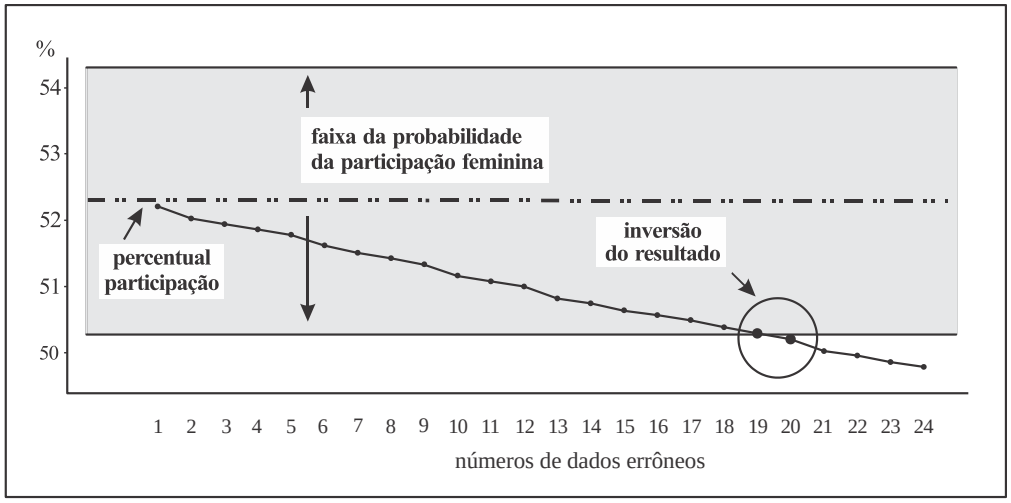

Figura 3: Dados errôneos e erro amostral
Fonte: elaborado pelos autores.

No caso da variável dicotômica "*indicação do posicionamento em relação ao sistema político"* com os atributos "*input*" e código dado "*1*" e "*output*" com o código dado "*2*", a inversão ocorre depois de 12 dados artificialmente digitados erroneamente ou 2,29 % dos dados válidos. No caso da variável múltipla "*religião*" com os atributos "*1*" para o atributo "*evangélico*", "*2*" para "*católico*", "*3*" para "outra" e "*4*" para "*não tem*", a inversão ocorreu depois de 36 dados digitados erroneamente. A inversão do resultado depende do tamanho da amostra, do número dos atributos declarados e do número dos dados errôneos. Pequenos erros amostrais facilitam a identificação, mas erros amostrais maiores a dificultam.

Há a possibilidade de que dados errôneos se compensem. No caso da pergunta "*Você tem um título de eleitor*?", 524 respostas com "*sim*" foram codificadas com "*1*" (65,5 %), 272 respostas "*não*" com "*2*" (34,5 %). Para este experimento, foram digitadas sete respostas

intencionalmente de forma errônea, sendo quatro com "*1*" ("*sim*"), embora devessem ter sido digitadas com "*2*" ("*não*"), e três com "*2*", embora a correta codificação fosse "*1*". Assim sendo, os dados errôneos se compensaram parcialmente e o novo resultado de 34,1 % que tivessem respondido com "*não*" representa uma diferença muito insignificante em relação ao valor verdadeiro, quando não se consideram possíveis vieses amostrais.

Para a pesquisa alcançar certo grau de confiabilidade, dados errôneos devem ser eliminados ou corrigidos. A mensuração dos conceitos escolares nas disciplinas Matemática e Português numa escala de 0 a 10 levou a 697 respostas válidas na disciplina em Matemática, com uma média de conceito de 7,52, e 699 respostas válidas em Português, com uma média de conceito de 7,36 wZquadro 4). Alguns alunos entrevistados indicaram como conceito os valores "*1*", "*2*", "*3*", "*4*", "*5*" e "*6*", os quais teriam causado uma reprovação na disciplina e seriam interpretados como dados suspeitos (1) (quadro 4). Dessa forma, estes valores foram eliminados da base dos dados, o que causou um aumento considerável da média das variáveis (2). Pensou-se também na substituição destes valores pela média geral de cada disciplina (3). Numa última substituição, os dados suspeitos foram substituídos pela média dos conceitos "*7*", "*8*", "*9*" e "*10*", valores que não tivessem levados à reprovação (4). Esse procedimento de exclusão e inserção de dados para fins de reparo (*constraint repair*) (Embury, Brand, Robinson, Sutherland, Bisby, Gray, Jones, & White, 2001, p. 675) ou manipulação destes (Bohannon, 2005, p. 143) pode causar efeitos variados (figura 4).

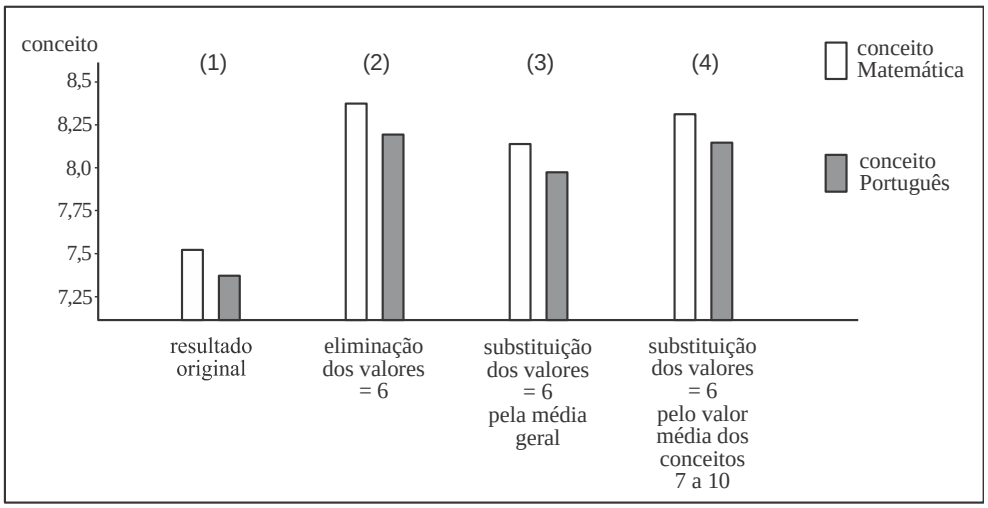

Figura 4: Eliminação e substituição de dados suspeitos de erro
Fonte: elaborado pelos autores.

5. CONCLUSÃO

A digitação manual deve ser, no médio prazo, o método mais usado de aquisição e informatização dos dados. Essa técnica não pode ser totalmente automatizada sem que ocorram perdas da qualidade. As etapas do processo de integração devem receber suporte de diferentes métodos para prevenir a aquisição de dados errôneos. O grande problema de dados errôneos na base de dados é que podem causar inversão dos resultados, apresentar informações de forma oposta, ocasionar dúvida acerca da geração de conhecimento e questionamentos sobre a confiabilidade dos métodos aplicados, bem como levar a uma validação errada.

Alguns procedimentos de prevenção de dados errôneos dependem não somente do aspecto humano, mas também do custo-benefício e rapidez. Cada etapa do processo de digitação de dados, como reconhecimento da pergunta e respectiva variável, decodificação da resposta, escolha do respectivo código, identificação do campo na planilha computacional, procura da tecla e digitação do dado, está sujeita a um processo sensorial−visual−cognitivo−táctil e representa, isoladamente ou em conjunto, uma fonte de erro. O reconhecimento de diferentes categorias qualitativas de um objeto abstrato se alcança pela codificação por meio de vários codificadores, o que indica mais custo e tempo, porém mais

confiança na validação dos dados. A dupla digitação dos dados também está sujeita a esta dicotomia custo-benefício. Meios técnicos, como o uso de sistemas de controle de entrada de dados, conseguem reduzir, mas não prevenir a digitação com erros. O reconhecimento acústico da digitação de certa tecla para respectivo dado é uma opção vantajosa, mas neste sistema sonoro-táctil a frequência destes sinais acústicos deve aumentar e acompanhar a velocidade táctil do digitador.

Uma base de dados absolutamente sem dados errôneos se alcança somente em casos específicos, o que demonstra a dificuldade de estabelecer uma definição precisa do termo "qualidade de dados" e de indicar métodos praticáveis para prevenir a aquisição de dados errôneos. O termo "qualidade" não é definido somente como uma representação adequada da base de dados para o respectivo fim de aplicação (*fitness for use*) (Bishop & Hank, 2018), mas precisa de indicadores numéricos para estabelecer um limite máximo de dados errôneos, a fim de determinar até que grau a base de dados deve estar livre de erros, ou seja, distingui-la em baixa, média ou alta qualidade. Um controle total não é racional em relação ao custo-benefício aceitável. A identificação de dados errôneos por meio de amostragem e de comparação dos dados digitados com os dados na fonte original representa o método mais apropriado. A substituição ou a modificação dos dados errôneos por dados mais apropriados, determinados por regras computacionais algorítmicas, caracterizam os dados como manipulados. Treinamentos intensivos, maior número de digitadores e incorporação de intervalos de tempo ajudam na prevenção de dados errôneos.

REFERÊNCIAS

Azeroual, O., Saake, G., & Abuosba, M. (2019). ETL best practices for data quality checks in RIS Databases. Informatics, 6(10), 1-13. https://doi:10.3390/informatics6010010

Bishop, B. W., & Hank, C. (2018). Measuring FAIR principles to inform fitness for use. International Journal of Digital Curation, 13(1), 35-46. https://doi10.2218/ijdc.v13i1.630

Bohannon, P., Fan, W., Flaster, M., & Rastogi, R. (2005, junho). A cost-based model and effective heuristic for repairing constraints by value modification. Proceedings of the ACM International Conference on Management of Data (SIGMOD), p. 143-154, Baltimore, MD, USA.

Braga, F., Lima, E. E. C., Leiva, G. de C., & Nascimento, A. G. de O. (2008, setembro). Fontes de dados populacionais no mundo: uma análise do relatório das Nações Unidas. Proceedings of the Congreso de la Asociación Latinoamericana de Población (ALAP), p. 1-8, Córdoba, Argentina, 3.

Broeck, J. v. d., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. Plos medicine, 2(10). Disponível em: <http://dx.doi.org/10.1371/journal.pmed.0020267>. Acesso em 14 dez. 2016.

Boeschoten, L., Oberski, D., & Waal, T. de (2017). Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (MILC). Journal of Official Statistics, 33(4), 921–962. https://doi.org/10.1515/JOS-2017-0044

Brislinger, E., & Moschner, M. (2019). Datenaufbereitung und Dokumentation. In U. Jensen, S. Netscher, & K. Weller (Eds.). Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten (pp. 97-114). Berlin: Verlag Barbara Budrich. https://doi.org/10.3224/84742233)

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: a survey. Mobile Networks and Applications, 19(2), 171-209. https://doi.org/10.1007/s11036-013-0489-0

Cohen, J. (1960) A coefficient for agreement of nominal scales. Educational and Psychological Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104

Cunha, M., Bastos, V. H., Veiga, H., Cagy, M., Mcdowell, K., Furtado, V., Piedade, R., & Ribeiro, P. (2004). Alterações na distribuição de potência cortical em função da consolidação da memória no aprendizado de datilografia. Arquivos de Neuro-Psiquiatria, 62(3-A), 662-668. https://doi.org/10.1590/S0004-282X2004000400018

Zio, M. d., Fursova, N., Gelsema, T., Gießing, S., Guarnera, U., Petrauskienė, J., Quensel von Kalben, L., Scanu, M., Bosch, K.O.T., Loo, M. v. d., & Walsdorfer, K. (2016). Methodology for data validation 1.0. Essnet Validat Foundation. Disponível em: <https://translateyar.ir/wp-content/uploads/2019/05/Methodology-for-data-validation-1.0.pdf>. Acesso em 13 jul 2020.

Embury, S. M., Brand, S. M., Robinson, J. S., Sutherland, I., Bisby, F. A., Gray, W. A., Jones, A. C., & White, R. J. (2001). Adapting integrity enforcement techniques for data reconciliation. Information Systems, 26(8), 657-689. https://doi.org/10.1016/S0306-4379(01)00044-8

Faulbaum, F. (2014). Total survey error. In N. Baur, & J. Blasius, (Eds.). Handbuch Methoden der empirischen Sozialforschung (pp. 439-456). Berlin: Springer.

Fowler, F. J. Jr. (2008). Survey research methods. Thousand Oaks: SAGE.

González-Prieto, A., Perez, j., Diaz, J., & López-Fernández, D. (2020). Inter-coder agreement for improving reliability in software engineering qualitative research. Disponível em: https://arxiv.org/pdf/2008.00977.pdf. Acesso em 05 abril 2020.

Graber, D. A. (2004). Methodological developments in political communication research. In L. L. Kaid (Ed.). Handbook of political communication research (pp.45-68). Mahwah: Lawrence Erlbaum Associates.

Granquist, L. (2011). Improving the traditional editing process. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. M. J. Colledge, & P. S. Kott (Eds.). Business Survey Methods (pp. 385-402). New York: John Wiley & Sons.

Hampel, F., Ronchetti, E., Rousseeuw, P., & Stahel, D. W. (1986). Robust Statistics. New York: John Wiley & Sons.

Hara, S., Nitanda, A., & Maehara, T. (2019). Data cleansing for models trained with SGD. Proceedings of the Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 33.

Henkel, K., & Almeida, J. de. (2003). Pesquisa quantitativa e de opinião pública sobre o ensino superior. Belém: UFPA.

Henkel, K. (2016). A validação de surveys políticos. A aplicação de teste paralelo, re-teste e avaliação externa em amostras como métodos da validação. Belém: UFPA.

Henkel, K. (2017). A categorização e a validação das respostas abertas em surveys políticos. Opinião Pública, 23(3), 786-808. https://doi.org/10.1590/1807-01912017233786

Hwang, C., Kim, H., & Jung, H. (2018). Detection and correction method of erroneous data using quantile pattern and LSTM. Journal of Information and Communication Convergence Engineering, 16(4), 242-247. https://doi.org/10.6109/jicce.2018.16.4.242

Krippendorff, K. (2004). Content analysis. Thousand Oaks: Sage.

Lau, S. H. (2018). Stress detection for keystroke dynamics. Dissertação de Mestrado, Universidade Carnegie Mellon, Pittsburgh, PA, USA

Lavalle, A., Maté, A., & Trujillo, J. (2020, março). An approach to automatically detect and visualize bias in data analytic. Proceedings of the International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP) e 23rd International Conference on Extending Database Technology, p. 84-88, Copenhagen, Dinamarca, 22.

Lindsey, D. R. B., & Logan, G. D. (2019). Item-to-item associations in typing: evidence from spin list sequence learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 45(3), 397-416. https://dx.doi.org/10.1037/xlm0000605

Liao, J., You, J., & Zhang, Q. (2019, abril). Research on library big data cleaning system based on big data decision analysis needs. Proceedings of the International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019), p. 377-382, Dalian, China, 3. https://doi.org/10.2991/icmeit-19.2019.62

Liua, S., Andrienko, G., Wu, Y., Cao, N., Jianga, L., Shi, C., Wang, Y. S., & Hong S. (2018). Steering data quality with visual analytics: the complexity challenge. Visual Informatics, 2(4), 191-197. https://doi.org/10.1016/j.visinf.2018.12.001

Maletic, J. I., & Marcus, A. (2000, outubro). Data cleansing. Beyond integrity analysis. Proceedings of the Conference on Information Quality IQ 2000, p. 200-209, Cambridge, MA, USA, 5.

Manrique-Vallier, D., & Reiter, J. P. (2017). Bayesian simultaneous edit and imputation for multivariate categorical data. Journal of the American Statistical Association, 112(520), 1708–1719. https://doi.org/10.1080/01621459.2016.1231612

Marsh, R. (2005). Drowning in dirty data? It's time to sink or swim: a four-stage methodology for total data quality management. Database Marketing & Customer Strategy Management, 12(2), 105-112. https://doi.org/10.1057/palgrave.dbm.3240247

Medeiros, M. (2005). Questionários: recomendações para formatação. Brasília: IPEA.

Möhring, W. & Schlütz, D. (2019). Die Befragung in der Medien- und Kommunikationswissenschaft. In W. Möhring, & D. Schlütz (Eds.). Das Interview als soziale Situation (pp. 41-67). Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-25865-8_2

Müller, H., Weis, M., Bleiholder, J., & Leser, U. (2005). Erkennen und Bereinigen von Datenfehlern in naturwissenschaftlichen Daten. Datenbank-Spektrum, 15, 26-35.

Rabin, E., & Gordon, A. M. (2004). Tactile feedback contributes to consistency of finger movement during typing. Experimental Brain Research, 155(3), 362-369. https://doi-org.ez3.periodicos.capes.gov.br/10.1007/s00221-003-1736-6

Rahm, E., & Do, H. H. (2000). Data cleaning: problems and current approaches. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 23(4), 3-13.

Rohwer, G. (2014). Deskriptive e funktionale Modelle in der statistischen Sozialforschung. In N. Braun, & N. J. Saam (Orgs.). Handbuch Modellbildung und Simulation in den Sozialwissenschaften (pp. 309-331). Berlin: Springer.

Saam, N. J. (2015). Einführung: Modellbildung und Simulation. In N. Braun, & N. Saam (Orgs.). Handbuch Modellbildung und Simulation in den Sozialwissenschaften (pp. 3-14). Berlin: Springer.

Saha, B., & Srivastava, D. (2014, abril). Data quality: the other face of big data. Proceedings of the IEEE International Conference on Data Engineering, p. 1294-1297, Chicago, IL, USA, 30.

Scheiner, J., Sicks, K., & Holz-Rau, C. (2014). Generationsübergreifende Mobilitätsbiografien – Dokumentation der Datengrundlage. Dortmund: Universität Dortmund.

Schomburg, H. (2001). Handbuch zur Durchführung von Absolventenstudien. Kassel: Universität Gesamthochschule Kassel.

Schwarz, H. (2018). Data consistency. In S. Netscher, & C. Eder (Eds.). Data processing and documentation: generating high quality research data in quantitative Social Science Research (pp. 25-33). Köln: GESIS.

Seligman, L., Rosenthal, A., Lehner, P., & Smith, A. (2002). Data integration: where does the time go. The Bulletin of the Technical Committee on Data Engineering, 25(3), 3-10.

Silva, E. D. da. (2013). Estudo da precipitação no Estado de Minas Gerais-MG. Dissertação de Mestrado, Universidade Federal de Itajuba, Minas Gerais, MG, Brasil.

Tröger P. (2018). Bedrohungen der Verlässlichkeit. In P. Tröger (Ed.). Unsicherheit und Uneindeutigkeit in Verlässlichkeitsmodellen. Wiesbaden: Springer Vieweg (pp. 83-123). https://doi.org/10.1007/978-3-658-23341-9_5

Waal, A. G. de. (2003). Processing of erroneous and unsafe data. Rotterdam: Universidade de Erasmus.

Winkler, W. E., & Chen, B. C. (2001, agosto). Extending the Fellegi-Holt model of statistical data editing. Proceedings of the Annual Meeting of the American Statistical Association, Survey Research Methods Section, Indianapolis, IN, USA.

Yip, C. (2007). Review Section: The production of knowledge: the challenge of social science research William H. Starbuck. New York: Oxford University Press, 2006. Management Learning, 38(3), 367-371. https://doi.org/10.1177/13505076070380030804