A QUALIDADE DE INDICADORES SOCIOECONÔMICOS PRODUZIDOS A PARTIR DE *BIG DATA*[1]

Rafael Bassegio Caumo[2]
João Artur de Souza[3]

A produção tradicional de indicadores socioeconômicos, predominantemente atrelada aos produtores de estatísticas oficiais e públicas, tem a aferição da sua qualidade consolidada ao longo das últimas décadas com diversas publicações relativas a *frameworks* e manuais de orientações ou boas práticas. Entretanto, o fenômeno da Revolução dos Dados vem colocando uma série de desafios aos produtores ao trazer o *big data* como possibilidade de fonte de dados a servir de insumo para os processos de produção. Entre os desafios está o de como garantir a qualidade dos produtos estatísticos. Nesse contexto, este artigo se propõe a resumir a produção técnico-científica sobre a qualidade de indicadores socioeconômicos produzidos a partir de *big data*. Para tanto, é conduzida uma revisão sistemática integrativa. Como resultados, o artigo apresenta uma percepção sobre: quais os principais agentes envolvidos com o tema; quais dimensões e *frameworks* de qualidade têm sido considerados e propostos; e quais as lacunas de pesquisa – direcionando esforços e trabalhos futuros.

Palavras-chave: indicadores socioeconômicos; estatísticas públicas; Revolução dos Dados; *big data*; qualidade.

THE QUALITY OF SOCIOECONOMIC INDICATORS PRODUCED FROM BIG DATA

The traditional production of socioeconomic indicators, mainly related to producers of official and public statistics, has its quality measured over the last decades according to various publications on frameworks, guidelines and manuals of good practices. However, the Data Revolution phenomenon has posed a series of challenges for producers when bringing big data as a source of data to serve as input for production processes – and among the challenges, how to guarantee quality. In this context, the present article proposes to synthetize scientific and technical production on the quality of socioeconomic indicators produced from big data. Therefore, a systematic integrative review is conducted. As results, the article presents a perception about: the main agents involved with the subject; what dimensions and quality frameworks have been considered and proposed; and what the research gaps are – directing efforts and future work.

Keywords: socioeconomic indicators; official statistics; Data Revolution; big data; quality.

1. DOI: http://dx.doi.org/10.38116/ppp57art9

2. Professor no Centro de Ciências da Administração e Socioeconômicas (Esag) da Universidade do Estado de Santa Catarina (Udesc). *E-mail*: <rbcaumo@gmail.com>.

3. Professor do Departamento de Engenharia do Conhecimento da Universidade Federal de Santa Catarina (UFSC). *E-mail*: <joao.artur@ufsc.br>.

LA CALIDAD DE INDICADORES SOCIOECONÓMICOS PRODUCIDOS A PARTIR DE *BIG DATA*

La producción tradicional de indicadores socioeconómicos, predominantemente vinculada a los productores de estadísticas oficiales y públicas, ha medido su calidad en las últimas décadas con base en diversas publicaciones sobre marcos, manuales y directrices de buenas prácticas. Sin embargo, el fenómeno de la Revolución de Datos ha planteado una serie de desafíos para los productores al llevar a *big data* como una fuente de datos que sirve como insumo para los procesos de producción, y entre los desafíos, cómo garantizar la calidad. En este contexto, el presente artículo propone resumir la producción científica sobre la calidad de los indicadores socioeconómicos producidos a partir de *big data*. Por lo tanto, se lleva a cabo una revisión integradora sistemática. Como resultado, el artículo presenta una percepción acerca de: los principales agentes involucrados con el tema; qué dimensiones y marcos de calidad han sido considerados y propuestos; y cuáles son los vacíos de investigación, dirigiendo los esfuerzos y el trabajo futuro.

Palabras clave: indicadores socioeconómicos; estadísticas públicas; Revolución de los Datos; *big data*; calidad.

JEL: C13; C18; C81; C83; O33.

1 INTRODUÇÃO

O fenômeno de explosão na quantidade de dados digitais sendo gerados e passíveis de armazenamento e processamento, produto da popularização da internet, do avanço tecnológico e da consolidação da era digital (Schmidt e Cohen, 2013), tem colocado uma oportunidade para agentes que buscam a descoberta e a aquisição de conhecimento sobre os mais diversos temas de interesse. Tal oportunidade permeia as fontes de *big data*, produtos da Revolução dos Dados (Kitchin, 2014), correspondentes a fontes de dados oriundos de rastros da interação humana com dispositivos digitais, tais quais: sensores, *scanners* e internet das coisas; *web* (páginas, redes sociais, mecanismos de buscas); aplicativos; telefones móveis; *global positioning systems* (GPSs); transações eletrônicas; registros administrativos entre outros.

São dados que vêm sendo percebidos como uma nova classe de ativos, ou como "o novo petróleo" (WEF, 2011; Bossoi, 2014), tamanha sua interessante relação de custo-benefício para processo de construção de conhecimento. Como consequência, desempenham papel central em um dito período de *data economy* (Opher *et al.*, 2016) e de quarta Revolução Industrial (Schwab, 2016) – impulsionados pela transformação digital e pela inteligência artificial. Fora do ambiente dos negócios, sua a relevância científica e de suporte socioeconômico também tem sido percebida (Asquer, 2013).

Os produtores de estatísticas públicas e/ou oficiais – disponibilizadas na maioria das vezes sob a forma de indicadores socioeconômicos de suporte a tomadas de decisão por parte tanto de gestores públicos, em seus processos de formulação e monitoramento de políticas públicas, quanto de agentes do setor privado, em

busca de competitividade para seus negócios – também estão atentos para esse fenômeno (UN, 2014), avaliando a dimensão da oportunidade e do ônus e bônus da incorporação das novas fontes. E isso não se dá por acaso, mas pelo fato de que muitos de seus processos de produção de indicadores são caros, demorados e de operacionalização complexa – uma vez que ainda estão fortemente concentrados em métodos tradicionais que envolvem pesquisas amostrais ou censos via levantamentos em campo com coleta conduzida por intermediários humanos.

Nesse contexto, a utilização de *big data* na produção de indicadores socioeconômicos pode trazer diversos benefícios (Citro, 2014), mas ainda depende de esforços de pesquisa que auxiliem no contorno a uma série de dificuldades que têm se colocado (Hackl, 2016), relativas a: capacidade organizacional (financeira, tecnológica e humana); cultura institucional; operacionalização e metodologia (captura, preparação, qualidade estatística e entrega); regulação e legislação; e redes de relacionamento externo.

No âmbito da justificativa considerada, esta proposta de artigo traz como problema de pesquisa a indagação sobre como garantir a qualidade de produtos estatísticos – incluindo indicadores socioeconômicos – construídos a partir de *big data*, em linha com a lacuna identificada por Hajnovic (2018, p. 1), ao afirmar que, "por ser um campo emergente, há pouca orientação para a mensuração da qualidade das aplicações de *big data* e *data science* em estatísticas oficiais".

Nesse contexto, o objetivo central consiste em resumir a produção técnico-científica sobre o tema por meio de uma revisão sistemática integrativa (Ercole, Melo e Alcoforado, 2014), identificando também dimensões e *frameworks* que têm sido considerados e propostos quando da aferição da qualidade de produtos estatísticos gerados a partir de fontes de dados digitais, resumidos por *big data* – produtos da Revolução dos Dados. Como resultado, espera-se que o levantamento permita identificar – além das dimensões e *frameworks* – tanto os principais agentes envolvidos com o tema quanto as lacunas de pesquisa atuais, direcionando esforços e trabalhos futuros.

Para tanto, este artigo é composto por, além desta seção introdutória: uma seção que traz um referencial teórico acerca dos principais conceitos aqui trabalhados (seção 2); outra com os procedimentos metodológicos utilizados para o atingimento do objetivo (seção 3); outra com resultados em termos de resumo bibliométrico, dimensões e *frameworks* de aferição de qualidade em estatísticas geradas a partir de *big data*, acompanhados da identificação de lacunas de pesquisa (seção 4); e, por fim, uma seção de conclusão que tece considerações finais (seção 5).

2 REFERENCIAL TEÓRICO

2.1 Indicadores socioeconômicos e estatísticas públicas

Um indicador é, geralmente, uma medida estatística quantitativa utilizada para ilustrar e comunicar um conjunto de fenômenos complexos de uma forma simples (OCDE, 2002; Bossel, 1999). Quando tratam de aspectos econômicos, demográficos, sociais e ambientais de uma nação, são ditos indicadores socioeconômicos, insumos fundamentais para o planejamento e a formulação de políticas e estratégias no mundo contemporâneo (Januzzi e Gracioso, 2002; Schnorr-Backer, 2016; Januzzi, 2002).

A produção de indicadores socioeconômicos é fortemente associada à produção de estatísticas oficiais e/ou públicas, uma vez que uma estatística poderá ser considerada ou utilizada para a construção de um indicador quando for ampla e reconhecidamente eficaz para descrever determinado aspecto da realidade. Ou seja, é no valor contextual baseado em uma teoria social ou finalidade programática que os indicadores socioeconômicos se diferem de simples estatísticas públicas. Assim, estatísticas oficiais e/ou públicas são apenas dados parcialmente preparados para uso na interpretação empírica da realidade, consistindo em insumos para a construção de indicadores socioeconômicos (Januzzi, 2001). Entre estatísticas oficiais e públicas, o diferencial se dá pelo fato de que as primeiras são chanceladas em nome do Estado ou organizações internacionais tidas como oficiais, e as segundas apenas carregam os mesmos princípios e objetivos, sem carimbo oficial (Schwartzman, 1997; IBGE, 2013).

Os agentes produtores dessas estatísticas estão predominantemente ligados ao setor público, aos institutos de pesquisa, à academia e ao terceiro setor, em virtude da necessidade de que os produtos sejam imparciais, confiáveis, de qualidade, respeitem padrões que permitam comparabilidade ao longo do tempo e entre diferentes localidades, entre outros princípios fundamentais (Unece, 1992; IBGE, 2013). Afinal, em sendo os produtores de estatísticas públicas encarregados de produzir e disponibilizar estatísticas sobre uma série de domínios e em diversos níveis de escala e granularidade (Kitchin, 2015), servindo a todos os setores da sociedade, faz-se necessário que se baseiem nas melhores práticas e em critérios de qualidade orientados por estatísticos com independência profissional e objetividade (Eurostat, 2014b).

Para tanto, o processo de produção de estatísticas públicas tem sido tradicionalmente conduzido à luz do rigor científico – por meio de métodos de pesquisa consolidados junto à comunidade –, utilizando-se, em geral, de práticas que envolvem pesquisas amostrais e censos – e, mais recentemente, valendo-se de registros administrativos (Hackl, 2016).

As pesquisas amostrais e os censos, práticas até então predominantes, são levantamentos que possuem utilidade e qualidade reconhecida por conta do sustento metodológico estatístico cientificamente consolidado, rigorosamente desenvolvido e padronizado ao longo de décadas (Kitchin, 2015). São métodos sustentados pela teoria da probabilidade (Cochran, 1977), que permitem alto potencial analítico ao viabilizarem que covariáveis sejam investigadas em conjunto com as variáveis principais de interesse ao mesmo tempo que podem ser controlados pelo pesquisador (Citro, 2014). Possibilitam que boa parte das fontes de erro associadas aos produtos sejam conhecidas – já estudadas, conforme sintetizam Brackstone (1999) e Biemer *et al*. (2014) –, facilitando mensurações relativas à qualidade e viabilizando processos inferenciais adequados para populações de interesse via tratamentos corretivos e validações estatísticas.

2.2 *Frameworks* tradicionais de qualidade

O conceito de qualidade aqui considerado parte da definição da *International Organization for Standardization* (ISO), que diz que qualidade é a totalidade de atributos e características de um produto ou serviço que culmina na sua capacidade de satisfazer necessidades explícitas ou implícitas (ABNT, 2005). Órgãos estatísticos utilizaram-se desta para definir o que viria a ser a qualidade de produtos estatísticos, ou, simplesmente, a qualidade estatística. FAO (2014) e Eurostat (2017) compartilham que se trata do grau em que produtos estatísticos atendem a requisitos em algumas dimensões de avaliação. Em linha, ONS (2013) diz que a qualidade de um produto estatístico pode ser definida como sua adequação ao seu propósito de utilização, ou sua capacidade de atender às necessidades dos usuários. Para UN (2018), a qualidade se faz pela combinação de todos os aspectos associados a quão bem processos e produtos estatísticos satisfazem as expectativas de usuários e outros *stakeholders*, caracterizando produtos adequados à demanda dos usuários e produzidos por meio de processos bem estruturados.

Dessas definições, pode-se perceber que a qualidade estatística é fruto da combinação de uma série de atributos/características de um produto, aqui chamada de "dimensões de qualidade", que:

- variam conforme a importância percebida pela demanda, isto é, de acordo com o interesse – perspectivas, necessidades e prioridades – do usuário final, comprometendo produtores com a maior quantidade possível de dimensões de qualidade – para que os produtos tenham maior alcance e disseminação perante diferentes perfis de usuários (FAO, 2014);
- variam conforme o produto estatístico em questão, uma vez que algumas dimensões podem se aplicar menos ou mais, de acordo com as características das fontes de dados e do resultado final desejado; e

- são frequentemente subdividas em domínios, como fazem, por exemplo: i) Eurostat (2015), ao organizar tais dimensões em ditas de ambiente institucional, processos de produção estatística e produtos estatísticos; e ii) Agafitei *et al*. (2015), ao organizá-las em de entrada/insumos (*input*), processo (*process/throughput*) e saída (*output*).

Dessa forma, quando da aferição da qualidade de produtos estatísticos, *frameworks* já desenvolvidos diferenciam-se pelas dimensões e domínios que consideram. Cabe ressaltar que por *frameworks* estão sendo consideradas estruturas construídas com o propósito de identificar elementos e suas relações a fim de nortear análises, explicando os processos e prevendo os resultados (Carvalho, 2013). Ou seja, correspondem neste contexto a estruturas conceituais que identificam dimensões e domínios de qualidade – bem como seus relacionamentos – de produtos estatísticos a fim de nortear o processo de aferição da qualidade.

Em âmbito geral, a maioria dos sistemas de gestão ou *frameworks* para aferição da qualidade de produtos e serviços parte dos princípios do *Total Quality Management* – TQM (Crosby, 1979; Saebo, 2016). Com foco específico sobre dados e informações, sob a ótica da ciência da informação, Zhu *et al*. (2012) propuseram o *Total Data Quality Management* (TDQM). Outros modelos e *frameworks* também propostos e utilizados no âmbito da ciência da computação e da tecnologia da informação podem ser encontrados em Jaya *et al*. (2017). Para o caso de indicadores socioeconômicos, entretanto, em virtude das características dos produtos e da demanda, tais princípios não se fazem suficientes (Saebo, 2016).

Características e valores desejados em indicadores socioeconômicos associam-se com os Princípios Fundamentais das Estatísticas Oficiais (Unece, 1992). Isso ocorre, primeiramente, pelo fato de que indicadores socioeconômicos são produtos estatísticos majoritariamente construídos a partir de estatísticas oficiais e públicas. Em segundo lugar, ocorre devido ao fato de que produtores de estatísticas oficiais são, por princípio, organizações altamente engajadas do ponto de vista do rigor técnico-científico considerado. Assim, a aferição da qualidade dos indicadores produzidos pode ser conduzida por *frameworks* propostos à luz do trabalho de Unece (1992).

Nesse contexto, um marco corresponde ao lançamento do *European Statistics Code of Practice* (CoP), do Gabinete de Estatísticas da União Europeia (Eurostat), estabelecido à luz dos Princípios Fundamentais das Estatísticas Oficiais e do TQM, em 2005 – e revisado em 2017 (Eurostat, 2017), após atualização da legislação estatística que reforçou o sistema estatístico europeu ao fortalecer o papel dos institutos nacionais de estatística (Regulação 2015/759 EU). Seu lançamento original estimulou a construção de outros códigos de recomendação para boas práticas estatísticas, como o *Recommendation of the OECD Council on good statistical practice*, da Organização para a Cooperação e Desenvolvimento Econômico – OCDE

(2015), o *Code of Practice for Official Statistics*, da Autoridade de Estatísticas do Reino Unido – revisado recentemente (UK Statistics Authority, 2018) –, e, no Brasil, o Código de Boas Práticas das Estatísticas do Instituto Brasileiro de Geografia e Estatística – IBGE (2013).

A versão revisada do CoP contempla uma declaração de qualidade endereçada ao Sistema Estatístico Europeu (ESS) e um *framework* de orientação para as práticas estatísticas, o *Quality Assurance Framework* (QAF), documentado por Eurostat (2015). Conforme Saebo (2016), praticamente todos os países europeus conduzem sua produção de estatísticas oficiais por meio de uma abordagem sistemática de gestão da qualidade baseada nos princípios do CoP.

Além do Eurostat, com o ESS QAF, diversos agentes envolvidos ou preocupados com produção de estatísticas oficiais e públicas também já desenvolveram *frameworks* de qualidade. Entre alguns, podem-se citar:

- *National Quality Assurance Framework* (NQAF), da Organização das Nações Unidas – ONU (UN, 2012);
- *Data Quality Framework* (DQF), do *Bureau* de Estatística da Austrália – ABS (2009);
- *Statistics Canada Quality Guidelines* (SCQG), do Instituto de Estatística do Canadá (Statistics Canada, 2009);
- *Statistics Quality Framework* (SQF), do Banco Central Europeu – ECB (2008);
- *Data Quality Assessment Framework* (DQAF), do Fundo Monetário Internacional – FMI (2003);
- *Statistics Quality Assurance Framework* (SQAF), da Organização das Nações Unidas para a Alimentação e a Agricultura – FAO (2014); e
- Código de Boas Práticas das Estatísticas, do IBGE (2013).

A título de breve comparação, o quadro 1 apresenta domínios, dimensões e a quantidade de indicadores ou elementos considerados por cada um dos *frameworks* mencionados. Como ponto de partida para aprofundamento, sugere-se Eurostat (2015) e HIQA (2018), que exploram as dimensões trazidas pelo ESS QAF e analisam em profundidade uma série de propostas de *frameworks*. No Brasil, o Código de Boas Práticas das Estatísticas do IBGE (2013), apesar de soar como um documento orientador com diretrizes e recomendações para práticas estatísticas, possui também um caráter de *framework* para aferição da qualidade estatística. O código do IBGE considera três domínios, dezessete dimensões (8, 4 e 5 em cada domínio) e oitenta indicadores (39, 18 e 23 em cada domínio).

QUADRO 1
Comparativo entre *frameworks* de qualidade para estatísticas oficiais e públicas selecionados

Framework	Domínio	Dimensão	Indicadores ou elementos
ESS QAF v1.2 (Eurostat, 2015)	Ambiente institucional	Comprometimento com a qualidade; confidencialidade estatística; imparcialidade e objetividade.	18 indicadores (4, 6 e 8 em cada dimensão)
	Processos de produção estatística	Metodologia sólida; procedimentos estatísticos apropriados; cargas de pesquisa não excessiva sobre respondentes; custo-benefício.	26 indicadores (7, 9, 6 e 4 em cada dimensão)
	Produtos estatísticos	Relevância; acurácia e confiabilidade; pontualidade e oportunismo temporal; coerência e comparabilidade; acessibilidade e clareza.	23 indicadores (3, 3, 5, 5 e 7 em cada dimensão)
UN NQAF (UN, 2012)	Sistema estatístico nacional	Coordenação do sistema nacional de estatística; relacionamento entre usuários e produtores; padrões estatísticos.	20 indicadores (4, 4 e 12 em cada dimensão)
	Ambiente institucional	Independência profissional; imparcialidade e objetividade; transparência; confidencialidade estatística e segurança; comprometimento com a qualidade; adequação dos recursos.	56 indicadores (10, 7, 6, 12, 15 e 6 em cada dimensão)
	Processos de produção estatística	Metodologia sólida; custo-benefício; procedimentos estatísticos apropriados; cargas de pesquisa não excessivas sobre respondentes.	44 indicadores (13, 9, 14 e 8 em cada dimensão)
	Produtos estatísticos	Relevância; acurácia e confiabilidade; pontualidade e oportunismo temporal; acessibilidade e clareza; coerência e comparabilidade; metadados.	43 indicadores (9, 4, 13, 5, 6 e 6 em cada dimensão)
ABS DQF (ABS, 2009)	Não explícito	Ambiente institucional; relevância; oportunismo temporal; acurácia; coerência; interpretabilidade; acessibilidade.	29 indicadores (6, 7, 2, 6, 4, 2 e 2 em cada dimensão)
SCQG (Statistics Canada, 2009)	Não explícito	Oportunismo temporal; relevância; interpretabilidade; acurácia; coerência; acessibilidade.	Não explícitos
ECB SQF (ECB, 2008)	Ambiente institucional	Independência e prestação de contas; autoridade para coletas; imparcialidade e objetividade; confidencialidade estatística; coordenação e cooperação; recursos e eficiência.	26 elementos (5, 1, 10, 3, 4 e 3 em cada dimensão)
	Processos de produção estatística	Metodologia sólida e procedimentos estatísticos adequados; custo-benefício e carga não excessiva sobre respondentes.	9 elementos (6 e 3 em cada dimensão)
	Produtos estatísticos	Relevância e completude; acurácia e confiabilidade; consistência/coerência; pontualidade e oportunismo temporal; acessibilidade e clareza.	28 elementos (5, 5, 5, 8 e 3 em cada dimensão)
FMI DQAF (FMI, 2003)	Não explícito	Pré-requisitos de qualidade; garantia de integridade; metodologia sólida; acurácia e confiabilidade; facilidade de manutenção; acessibilidade.	51 indicadores (10, 8, 6, 10, 8 e 9 em cada dimensão)

(Continua)

(Continuação)

Framework	Domínio	Dimensão	Indicadores ou elementos
FAO SQAF (FAO, 2014)	Ambiente institucional	Independência profissional e imparcialidade; confidencialidade estatística; comprometimento com a qualidade; padrões internacionais; cooperação; coordenação com outros produtores.	31 elementos (8, 7, 6, 4, 3 e 3 em cada dimensão)
	Processos de produção estatística	Metodologia sólida e procedimentos estatísticos adequados; custo-benefício; carga não excessiva sobre respondentes.	21 elementos (11, 4 e 6 em cada dimensão)
	Produtos estatísticos	Relevância; acurácia e confiabilidade; pontualidade e oportunismo temporal; coerência e comparabilidade; acessibilidade e clareza.	21 elementos (4, 5, 4, 4 e 4 em cada dimensão)
Código de Boas Práticas das Estatísticas (IBGE, 2013)	Ambiente institucional e coordenação	Independência institucional; coordenação do sistema estatístico nacional; mandato estatístico de coleta de dados; confidencialidade estatística; uso eficiente dos recursos; compromisso com a qualidade; imparcialidade e objetividade; cooperação e participação internacional.	39 indicadores (7, 3, 4, 7, 4, 4, 7, 3 em cada dimensão)
	Processos estatísticos	Metodologia sólida; processos estatísticos adequados; solicitação de informação não excessiva; relação entre custo e eficácia.	18 indicadores (5, 7, 3, 3 em cada dimensão)
	Produtos estatísticos	Relevância; precisão e acurácia; oportunidade e pontualidade; coerência e comparabilidade; acessibilidade e transparência.	23 indicadores (4, 5, 5, 4, 5 em cada dimensão)

Elaboração dos autores.

Cabe destacar que, em relação à mensuração da acurácia, dimensão presente em todos os *frameworks* de qualidade apresentados, a abordagem de fundo predominantemente utilizada é a do *Total Survey Error* (TSE), com raízes descritas por Groves e Lyberg (2010), Biemer (2010) e Biemer *et al*. (2017). A abordagem do TSE se baseia no *total survey error paradigm* (Platek e Sarndal, 2001), que entende que as grandes fontes de erros de uma pesquisa devem ser identificadas de modo que esforços e recursos para minimizá-las possam ser direcionados. Tal paradigma é parte do conceito maior de *Total Survey Quality* (TSQ), que aborda a qualidade no contexto da adequação da estimativa produzida ao seu fim (Juran e Gryna, 1980).

O TSE considera a acurácia como o aspecto principal da qualidade de uma estimativa gerada por uma pesquisa do tipo *survey* – composta pelo viés e pela precisão (variância) –, medida pelo erro quadrático médio e impactada por fontes de erros dos tipos: amostrais, tais quais de desenho amostral, de tamanho amostral e de escolha do estimador; e não amostrais, como os de especificação, de não respostas, de cobertura, de mensuração e de processamento dos dados (Biemer, 2010; Biemer *et al.*, 2014). No âmbito de pesquisas mais complexas ou da utilização de dados que derivam de fontes diferentes de *surveys*, como os de registros administrativos (Daas *et al.*, 2009; Nederpelt, 2010; Nederpelt e Daas, 2012; Iwig *et al.*, 2013; Wallgren e Wallgren, 2014), importantes discussões são feitas por Biemer *et al*. (2017) e Reinert e Stoltze (2016).

É possível perceber que, dos códigos de recomendação e *frameworks* aqui apresentados, alguns tem por objetivo servir apenas de sugestão – ou ponto de partida genérico – para países quando da estruturação de seu sistema de produção estatística, identificando atividades, métodos e ferramentas que podem ser utilizadas. Entretanto, cabe ressaltar que, em geral, este tipo de documentação vem evoluindo: i) passando de "*assessement*" para "*assurance*", isto é, de suporte para adoção de práticas internas para ferramenta de certificação; ii) ampliando sua aplicabilidade daquela restrita a produtos estatísticos gerados a partir de pesquisas amostrais do tipo *survey* para uma utilizável em produtos gerados a partir de outras fontes – como registros administrativos (ONS, 2013); e iii) expandindo a aferição da qualidade dos produtos para a qualidade dos processos de produção como um todo, com sistemas e metodologias-padrão, impulsionados pela introdução do *Generic Statistical Business Process Model* (GSBPM) (Unece, 2013a) – que passou a servir de base para a grande maioria dos países europeus na estruturação de seus modelos de processo de negócio estatístico (Eurostat, 2016).

Um exemplo representativo dessa evolução é o *Guidelines for Measuring Statistical Output Quality – Version 4.1* (ONS, 2013), do Departamento Nacional de Estatísticas do Reino Unido. O documento fornece um *checklist* de medidas e indicadores para mensuração da qualidade estatística que engloba diversos aspectos

evolutivos mencionados. O manual sugere métricas para todas as etapas do GSBPM no âmbito da mensuração de cada uma das cinco dimensões de qualidade de produtos estatísticos consideradas pela ESS (relevância, acurácia e confiabilidade, pontualidade e oportunismo temporal, acessibilidade e clareza, coerência e comparabilidade), tanto para contextos de dados oriundos de pesquisas amostrais quanto de registros administrativos ou de múltiplas fontes.

Em suma, conforme observa Agafitei *et al.* (2015), a aferição da qualidade da produção tradicional de estatísticas públicas alcançou um estágio de maturidade que considera três domínios: domínio de insumos (*input*), associado a fontes produtoras e aos dados brutos; domínio de processos de produção (*process*), associado aos métodos e procedimentos empregados durante a geração dos produtos estatísticos; e domínio de produtos finais (*output*), associado a características que atendem aos interesses dos usuários. Os mesmos autores destacam que, apesar de a mensuração da qualidade dos domínios de *input* e *output* ser tradada de forma consensual perante os diversos mecanismos, a dimensão da qualidade do processo é algumas vezes avaliada sob a ótica do GSBPM e do TQM, com foco nas etapas do processo em si e objetivando melhorias de interesse do produtor, e outras vezes avaliada sob a ótica de como cada etapa do processo de produção influencia na qualidade do produto final, com foco na demanda – no que deseja o usuário final.

De um modo geral, o que fica é que *frameworks* de qualidade devem ser suficientemente genéricos para que abarquem diversos contextos, com importância dada a cada domínio e dimensão de qualidade decidida caso a caso (ECB, 2008) – com base nas características das fontes de dados e dos resultados finas desejados, necessidades da demanda.

Essa percepção também existe fora do escopo das estatísticas públicas e oficiais, conforme a revisão de literatura no âmbito da ciência da informação de Fagundes, Macedo e Freund (2017). Para Valente e Fujino (2016), mesmo que não exista consenso sobre o conceito de qualidade e até mesmo sobre o significado exato de cada dimensão (Batini *et al.*, 2009), diversas são as proposições de dimensões de qualidade de dados e informações, dependentes da abordagem e da vertente de cada aplicação. Tal percepção é carregada até os dias atuais, conforme corroboram Gudivada, Apon e Ding (2017) e Jaya *et al.* (2017), ao dizerem que a qualidade dos dados está altamente associada ao contexto de utilização final e sua sinergia, às necessidades dos clientes, habilidades de utilização e de acesso aos dados. Em linha, Wang e Strong (1996) propõem que a qualidade de dados seja considerada de maneira genérica, para que possa ser aplicada a diferentes domínios e contextos, classificando quinze dimensões de qualidade em quatro categorias: intrínsecas (associadas à acurácia); contextuais (associadas à relevância); representacionais (associadas à forma de representação) e de acessibilidade (associadas à forma de acesso).

Entretanto, fenômenos recentes vêm trazendo oportunidades e desafios relativos a reestruturação dos modelos tradicionais de produção estatística e, consequentemente, à aferição da qualidade correspondente: o *big data* e a Revolução dos Dados.

2.3 Revolução dos Dados e *big data*

Há algum tempo já se percebe uma avalanche de dados sendo criada diariamente (Miller, 2010; Helbing *et al*., 2016), impulsionada pela popularização da internet, pelo avanço das tecnologias da informação e pela consolidação da era digital (Schmidt e Cohen, 2013). Tais dados, ditos dados digitais, derivam dos rastros – tanto inconscientes quanto deliberados – da atividade humana e não humana capturada por dispositivos inseridos no universo dos sensores, *scanners* e internet das coisas, dos *smartfones* e aplicativos, das páginas da internet, dos mecanismos de buscas, das redes sociais virtuais, dos *blogs* e *web* fóruns, entre outros. Por vezes, o termo *big data* surge como sinônimo desses dados digitais, como faz Unece (2013b), ao apresentar e classificar as possíveis fontes desse tipo de dado.

Nesse contexto, se, no passado, a predominância por dados no mundo estava vinculada a dados ditos *small data*, predominantemente analógicos, escassos e de acesso limitado, hoje, a quase totalidade dos dados produzidos é digital, gerados em vasta escala, com velocidade e em variedade de domínios e formatos (Kitchin, 2015). Essa transformação compõe o processo do que é por Kitchin (2014) chamado de Revolução dos Dados.

Antes mesmo de ter seu conceito formalizado, esse processo já era percebido como tendo características de inovação de ruptura ao desafiar o *status quo* de como os dados são produzidos, geridos, analisados, armazenados e utilizados, modificando a maneira como o conhecimento é produzido, como os negócios são conduzidos e como a governança é promulgada. Há alguns anos, Mayer-Schönberger e Cukier (2013) já diziam que o *big data* se tratava um fenômeno que estaria pronto para "chacoalhar" tudo, dos negócios e ciências até os sistemas de saúde, os governos, a educação, a economia, as humanidades e todos os outros aspectos da sociedade. De fato, tem-se percebido que sim.

2.4 Oportunidades e desafios do *big data* para a produção de indicadores socioeconômicos

Os produtores de estatísticas públicas e/ou oficiais – disponibilizadas na maioria das vezes sob a forma de indicadores socioeconômicos – também já atentaram para oportunidades trazidas pelo *big data* e pela Revolução dos Dados (UNSC, 2014; Eurostat, 2013). Vislumbram possibilidades de contorno de uma série de problemáticas associadas ao modelo tradicional de produção de estatísticas, fortemente baseados em pesquisas amostrais e censos via levantamentos em campo, com coleta conduzida por intermediários humanos – processos, em geral, caros, demorados e de operacionalização complexa.

Fala-se em oportunidade para contornar problemáticas relativas a aspectos de: recursos – tempo e custo financeiro (Hackl, 2016; Daas e Puts, 2014; Kitchin, 2015; Tam e Clarke, 2014; Eurostat, 2014b); atendimento à demanda (Harwood e Mayer, 2016; Florescu *et al.*, 2014); operacionalização e logística (Citro, 2014; Struijs, Braaksma e Daas, 2014; Braaksma e Zeelenberg, 2015); metodologia (Manski, 2014; Hand, 2015); e posicionamento estratégico e sobrevivência para as instituições produtoras (Demunter, 2017; Struijs e De Broe, 2018; MacFeely, 2016).

Entretanto, a utilização de *big data* na produção de indicadores socioeconômicos e estatísticas públicas traz também incertezas e desafios, despertando a necessidade por esforços e investimentos em pesquisa e articulação política. Entre alguns já percebidos, podem-se citar aspectos de:

- operacionalização da produção e metodologia (Hassani, Saporta e Silva, 2014; Tennekes, De Jonge e Daas, 2011; Zikopoulos *et al.*, 2012; Fry, 2008; Daas *et al.*, 2012; Liu, Jiang e Heer, 2013; Couper, 2013; Daas *et al.*, 2013; Unece, 2015; Reimsbach-Kounatze, 2015; Braaksma e Zeelenberg, 2015; Citro, 2014; Tufekci, 2013; Rudin *et al.*, 2014; Buelens *et al.*, 2012; Flekova e Gurevych, 2013; Hastie, Tibshirani e Friedman, 2009; Daas, 2012; Daas e Puts, 2014; O'Connor *et al.*, 2010);
- capacidade organizacional – financeira, tecnológica e humana (Mills *et al.*, 2012; NAS, 2013; Schutt e O'Neil, 2013; Parise, Iyer e Vesset, 2012; Scannapieco, Virgillito e Zardetto, 2013; Struijs, Braaksma e Daas, 2014; Tam e Clarke, 2014; Cervera *et al.*, 2014; Dunne, 2013; Davenport e Patil, 2012; LaValle, 2011; UNGP, 2012; Eurostat 2014b);
- regulação, legislação, ética e privacidade (Wirthmann e Reis, 2018; Tam e Kim, 2018; Nasem, 2017a; Vaccari, 2014; Hackl, 2016; Vale, 2015; Struijs e Daas, 2013; Daas *et al.*, 2015; De Jonge, van Pelt e Roos, 2012);
- redes de relacionamento externo (ESSnet, 2013; Tam e Clarke, 2015a; Krätke e Byiers, 2014); e
- cultura institucional (Kitchin, 2015).

Entre os desafios, a qualidade de estatísticas e indicadores gerados a partir de fontes de dados do tipo *big data*, presente no aspecto metodológico-operacional, é o elemento de interesse deste artigo.

3 METODOLOGIA

Este estudo se propõe a ser uma pesquisa científica no âmbito das ciências empíricas sociais, conforme definições trazidas por Gil (2008). As bases lógicas para verificação e validação científica do conhecimento construído estão apoiadas no paradigma

positivista, originário dos trabalhos dos empiristas Bacon, Hobbes, Locke e Hume durante os séculos XVI, XVII e XVIII, posteriormente formalizado por Auguste Comte no século XIX (Trivinos, 1992). Dessa forma, entende-se a realidade do mundo como objetiva, composta por coisas e fatos. E, nesse ambiente, a pesquisa quantitativa surge com intuito de descobrir, pelo raciocínio e a observação, as leis que regem essa realidade, testando teorias objetivas e examinando a relação entre as variáveis (Creswell, 2010; Trivinos, 1992; Bryman, 2004).

Em relação ao alcance do objetivo de resumir a produção técnico-científica sobre a qualidade de produtos estatísticos gerados a partir de *big data*, este estudo almeja alcançar o nível exploratório – objetivando proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito, com delineamento do tipo bibliográfico –, sendo elaborado a partir de materiais já publicados (Gil, 2008).

O levantamento bibliográfico é integrativo (Ercole, Melo e Alcoforado, 2014) e realizado de forma sistemática (Forbes, 1998), partindo de uma pergunta claramente formulada e utilizando métodos sistemáticos e explícitos para: identificar, selecionar e avaliar criticamente pesquisas relevantes; e coletar e analisar dados dos estudos (Green e Higgins, 2005), com critérios para a busca que seguem o método Prisma (Moher *et al.*, 2009). Ao final, os documentos coletados serão analisados em busca de conteúdos que permitam responder ao objetivo central do artigo – com identificação dos principais agentes envolvidos, das dimensões e *frameworks* de qualidade e das lacunas de pesquisa.

Os documentos foram selecionados a partir de levantamento realizado em 23 de junho de 2018 que percorreu inicialmente as bases apresentadas no quadro 2. As bases *Scopus* e *Web of Science* foram escolhidas por sua reconhecida cobertura de literatura científica revisada (*peer-reviewed*), enquanto a Scielo foi selecionada por sua ampla cobertura de publicações em nível brasileiro.

QUADRO 2
Informações sobre a origem dos primeiros documentos selecionados

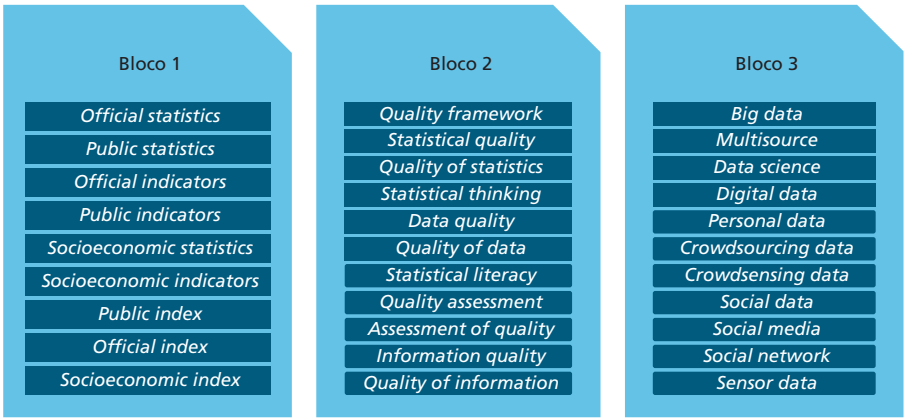
Base	Tipo de busca	Retornos
Scopus (Elsevier)	Título, resumo e palavras-chave	13 documentos
Web of Science (Clarivate Analytics)	Tópico	8 documentos
Scielo.org	Padrão	0 documento

Elaboração dos autores.

O critério de busca selecionou documentos que contivessem pelo menos uma das expressões de cada um dos três blocos apresentados na figura 1. Dessa forma, a busca percorreu 9 x 11 x 11 = 1.089 possíveis combinações das expressões contempladas pelos três blocos. Dos 21 documentos que retornaram às buscas nas

diferentes bases, seis eram repetidos e foram excluídos, fazendo com que a quantidade de documentos exclusivos considerados de tais bases fosse de quinze. Destes, sete foram excluídos em uma análise de conteúdo, ou por terem sido julgados fora do escopo de interesse, ou por estarem escritos em língua diferente de inglês ou português. Assim, apenas oito documentos derivados das bases consultadas foram aqui trabalhados.

FIGURA 1
Critérios de busca utilizados na primeira etapa do levantamento bibliográfico



Elaboração dos autores.

Em virtude da baixa produção científica encontrada no levantamento, foi realizada também uma busca por documentos de outros tipos – relatórios técnicos, textos para discussão etc. – no mecanismo de buscas na *web* do *Google* – tornando a busca integrativa (Ercole, Melo e Alcoforado, 2014). Nessa etapa, o critério de busca foi por resultados que contivessem pelo menos uma expressão de cada um dos três grupos: i) *official statistics*, *public statistics* ou *official indicators*; ii) *quality framework*, *statistical quality*, *data quality*, *quality assessment* ou *information quality*; e iii) *big data* ou *multisource*. Dos resultados, os cem primeiros foram percorridos e analisados para que se pudesse manualmente selecionar 46 alinhados com os objetivos deste estudo. Dessa forma, a revisão bibliográfica integrativa (Ercole, Melo e Alcoforado, 2014) considerou, ao final, 54 documentos. Uma vez que o algoritmo de classificação e ordenamento dos resultados utilizados pelo *Google* considera a experiência do usuário no que se refere a mais de duzentos fatores,[4] é de conhecimento dos autores deste artigo que os cem primeiros resultados não necessariamente serão os mesmos para diferentes usuários.

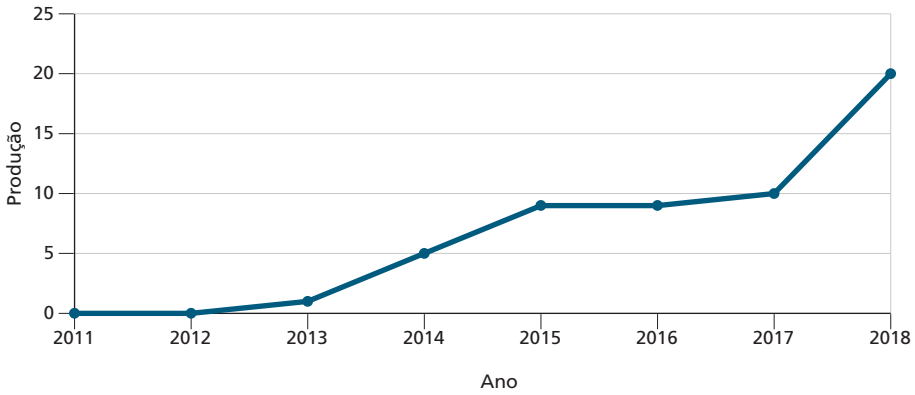
4. Como a Pesquisa Google funciona. Disponível em: <https://bit.ly/3uAUM0V>. Acesso em: 10 jul. 2019.

4 RESULTADOS

4.1 Resumo bibliométrico

A evolução temporal da produção absoluta dos materiais que compõem a base técnico-científica aqui analisada é apresentada no gráfico 1. Destaca-se a atualidade e o crescimento do tema, baseando-se no fato de que não foram encontradas publicações anteriores ao ano de 2013 e que a quantidade de produções não decresce desde então. Em 2018, mesmo que a busca tenha sido realizada no meio do ano, a quantidade de publicações encontradas foi impulsionada pela importância dada ao tema na conferência intitulada *European Conference on Quality in Official Statistics*, realizada em junho na Cracóvia, Polônia.

GRÁFICO 1
Evolução temporal da produção quantitativa dos documentos analisados, por ano (2011-2018)



Elaboração dos autores.
Obs.: O quantitativo de 2018 corresponde ao encontrado na data da busca.

Como forma de análise dos pesquisadores e instituições mais engajados com o assunto, foram filtrados todos os autores que apareciam em mais de um documento analisado. Os resultados, com nome do autor, quantidade de publicações em que aparece e país e nome da instituição de vínculo especificada nos documentos, estão apresentados no quadro 3. A produção é predominantemente europeia e, como poderia se esperar, concentrada em órgãos de estatística e de pesquisa do setor público e do terceiro setor. Cabe destacar também a ausência de documentos na base Scielo.org, conforme demonstra o quadro 2, sugerindo uma baixa ou inexistente produção sobre o tema em âmbito brasileiro.

QUADRO 3
Pesquisadores e quantidade de publicações, com país e nome da instituição de vínculo

Autor	Quantidade	País da instituição	Instituição
Brancato, G.	5	Continental (Europa); Itália	Eurostat; *Italian National Statistical Institute*
Reis, F.	3	Continental (Europa)	Eurostat
Struijs, P.	3	Holanda	*Statistics Netherlands*
Ascari, G.	2	Itália	*Italian National Statistical Institute*
Unece[1]	2	Continental (Europa)	*United Nations Economic Commission for Europe*
Clarke, F.	2	Austrália	*Australian Bureau of Statistics*
de Waal, T.	2	Continental (Europa); Holanda	Eurostat; *Statistics Netherlands*
di Consiglio, L.	2	Continental (Europa); Itália	Eurostat; *Italian National Statistical Institute*
Nasem[2]	2	Estados Unidos	*National Academies of Sciences, Engineering, and Medicine*
Radini, R.	2	Itália	*Italian National Statistical Institute*
Scannapieco, M.	2	Itália	*Italian National Statistical Institute*
Scholtus, S.	2	Continental (Europa); Holanda	Eurostat; *Statistics Netherlands*
Tam, S. M.	2	Austrália	*Australian Bureau of Statistics*
Uluwiyah, A.	2	Indonésia	*Statistics Indonesia*
Vaju, S.	2	Continental (Europa)	Eurostat
van Delden, A.	2	Continental (Europa); Holanda	Eurostat; *Statistics Netherlands*
Wirthmann, A.	2	Continental (Europa)	Eurostat

Elaboração dos autores.
Notas: [1] Unece – United Nations Economic Commission for Europe.
[2] Nasem – National Academies of Sciences, Engineering, and Medicine.

Por fim, a figura 2 apresenta, de maneira visual, as palavras-chave listadas nas publicações selecionadas, dimensionadas proporcionalmente à quantidade de aparições.

FIGURA 2
Nuvem de palavras-chave das publicações selecionadas

Elaboração dos autores.

4.2 Dimensões de qualidade no contexto de *big data*

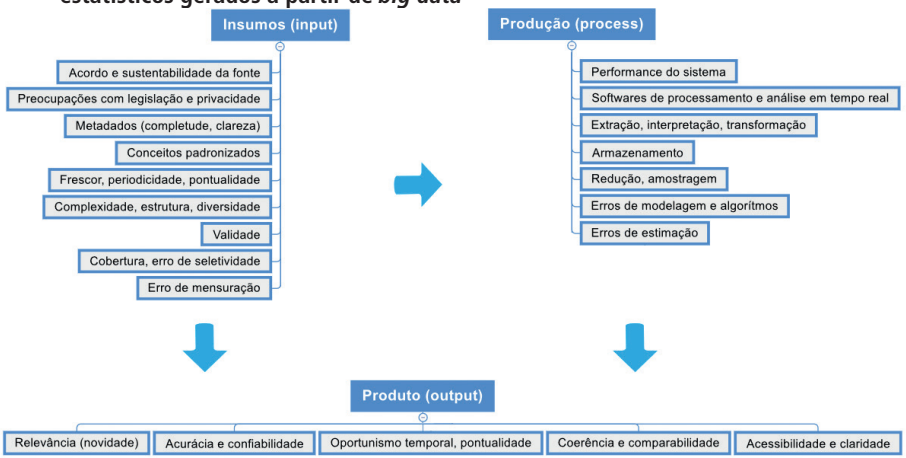
No âmbito da apresentação das dimensões de qualidade a serem consideradas quando da utilização de *big data* para a produção de estatísticas públicas, tomou-se como ponto de partida uma interessante revisão de literatura realizada por Brancato e Di Consiglio (2018). Os autores perceberam quatro correntes de pesquisa que abordaram a questão da qualidade sob as óticas: i) da extensão dos conceitos e métodos tradicionais de aferição da qualidade para a produção baseada em *big data* (Struijs e Daas, 2014; Unece, 2015; Kreuter *et al.*, 2015; Baker, 2017; Cai e Zhu, 2015); ii) da percepção de possíveis fontes de erros quando da tentativa de execução de aplicações práticas de produção de estatísticas públicas a partir de *big data* (Lugomer *et al.,* 2017; Hsieh e Murphy, 2017; ESSnet, 2018; Wang *et al.*, 2012; Barcaroli, Golini e Righi, 2018); iii) do grau de confiabilidade das fontes de *big data* e de usabilidade destes dados (Eurostat, 2014a); e iv) do processo de produção das estatísticas (Tao e Gao, 2016; Cai e Zhu, 2015).

A figura 3 apresenta as principais dimensões de qualidade identificadas por Brancato e Di Consiglio (2018), caracterizando-as entre associadas aos insumos e condições institucionais (*input*) ou associadas ao processo de produção de estatísticas (*process*), relacionando-as com as dimensões de qualidade de produtos

estatísticos (*output*) reconhecidas pelo ESS (Eurostat, 2015) – consideradas em diversos *frameworks* tradicionais.

Durante revisão sistemática aqui realizada, pôde-se perceber que diversos autores corroboram as dimensões reunidas por Brancato e Di Consiglio (2018). Nem todos percebem todas as dimensões ao mesmo tempo, concentrando-se em esferas específicas. Alguns autores, por sua vez, mencionaram dimensões não explicitamente contempladas no resumo da figura 3. A grande maioria, entretanto, possui alguma relação e/ou poderia ser facilmente incorporada – tratada como elemento ou subdimensão – às dimensões de qualidade elencadas na figura apresentada.

FIGURA 3
Relações entre dimensões de qualidade de insumos, processo de produção e produtos estatísticos gerados a partir de *big data*



Fonte: Traduzida de Brancato e Di Consiglio (2018).
Obs.: Figura cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Aspectos como mínima granularidade, mencionado por Nasem (2017b) e Gurwitz (2012), e usabilidade, citado por Unece (2015), podem ser parte da dimensão relevância. Processos de gestão e governança de dados, com integração, linkagem, limpeza, processamento, ponderação e imputação de dados (Vaju e Meszaros, 2018; Thompson, 2018; Máslankowski e Nowicka, 2018; Lauro e Traverso, 2018; Hand, 2018; Struijs e Daas, 2014; Nasem, 2017b; Biffignandi e Signorelli, 2015; Unece, 2015; Uluwiyah, 2016) podem fazer parte das dimensões extração, interpretação e transformação. Variabilidade, volatilidade, instabilidade e disrupção das fontes (Silva, 2016; Biffignandi e Signorelli, 2015; Hajnovic, 2018) entram em sustentabilidade da fonte, em linha com mudanças metodológicas que podem acontecer (Hand, 2018) e culminar na perda da comparabilidade temporal. Comunicação de incertezas (Hand, 2018) e interpretabilidade (ABS, 2009; Statistics

Canada, 2009; Hackl, 2016; Tam e Clarke, 2015b; Nasem, 2017b) entram em acessibilidade e clareza – possuindo relação também com os metadados. Veracidade (Bergamaschi *et al.*, 2016; Silva, 2016), consistência (Lauro e Traverso, 2018; Tam e Clarke, 2015b), reprodutibilidade (Struijs e De Broe, 2018) e auditabilidade (Cai e Zhu, 2015) são incorporadas tanto pela acurácia e confiabilidade quanto pela coerência e comparabilidade. Alta dimensionalidade, mencionada por Biffignandi e Signorelli (2015), pode ser contemplada pelas dimensões redução, modelagem e/ou estimação. Eficiência, *performance* e infraestrutura computacional (Al-Hajjar *et al.*, 2015; Tam e Clarke, 2015b; Uluwiyah, 2016), assim como escalabilidade e interoperabilidade (MacFeely, 2016; Hilbert, 2016), relacionam-se com as dimensões *performance* do sistema e *softwares* de processamento, tendo as últimas uma relação também com a coerência e a comparabilidade.

Por outro lado, alguns aspectos percebidos parecem não se relacionarem diretamente com as dimensões apresentadas na figura 3. Autores como Hackl (2016) e Tam e Clarke (2015b) entendem o componente custo como uma dimensão da qualidade. Também na linha dos recursos utilizados, Batini *et al*. (2015) defendem que, para serem de qualidade, os produtos devem atentar ao aspecto da redundância – com minimalidade, compactude e concisão. Os aspectos transparência, integridade, independência e ambiente da instituição produtora (Nasem, 2017b; Unece, 2015) também não estão explicitamente contemplados. E, por fim, a possibilidade de manipulação – no sentido pejorativo – dos dados de algumas fontes também deve ser levada em consideração, conforme Wirthmann e Reis (2018) e MacFeely (2016). A consideração de tais aspectos merece ser mais profundamente debatida pelos proponentes de modelos de aferição da qualidade de produtos estatísticos.

Para Agafitei *et al*. (2015), a mensuração da qualidade deve ser baseada exclusivamente nas dimensões do *output*, apesar de reconhecerem que as dimensões associadas aos insumos e aos processos de produção também influenciam na qualidade final de produtos estatísticos. Os autores argumentam que a mensuração de todas as dimensões nos três domínios – *input, process, output* – frequentemente não se faz viável no contexto de *big data*, especialmente quando diversas fontes de dados estão sendo integradas (*multisource*) durante a geração de um produto. Como bases de *big data* não são originalmente estruturadas para responder aos questionamentos inerentes à produção de estatísticas públicas, provavelmente precisarão ser combinadas com dados de outras fontes (pesquisas do tipo *survey*, registros administrativos ou até mesmo outras bases de *big data*), fazendo com que a complexidade dos contextos de múltiplas fontes (*multisource*) seja corriqueira (Reis *et al.*, 2016).

Ao defenderem sua proposta, os autores consideram as dimensões de qualidade de produto do ESS (Eurostat, 2015): i) relevância, associada ao atendimento das necessidades e expectativas dos usuários por parte do produto; ii) acurácia e confiabilidade, associadas à capacidade do produto de retratar a realidade de forma precisa, não enviesada e confiável; iii) temporalidade e pontualidade, associadas a uma entrega rápida e no momento que se fizer útil; iv) coerência e comparabilidade, associadas a utilização de conceitos padronizados amplamente aceitos, permitindo combinação e comparação com estatísticas de outras localidades e temporalidades; e v) acessibilidade e clareza, associadas a produtos entregues de maneira conveniente, em formatos claros e com informações relevantes completas e compreensíveis.

Destas, algumas podem ser aferidas em estatísticas produzidas a partir de *big data* da mesma forma que são em estatísticas tradicionalmente produzidas. As dimensões (i) relevância e (v) acessibilidade e clareza, por exemplo, podem ser igualmente percebidas e mensuradas em ambos os contextos – cabendo destacar que produtos gerados de *big data* provavelmente terão clareza prejudicada, uma vez que podem ser originários de bases privadas. As dimensões (iii) temporalidade e pontualidade são também igualmente percebidas e mensuradas em ambos os cenários, com expectativas de mais satisfatórias quando se trabalhando com *big data*. Assim, as dificuldades de mensuração da qualidade no âmbito dos produtos finais concentrar-se-ão nas dimensões (ii) acurácia e confiabilidade e (iv) coerência e comparabilidade, uma vez que são altamente impactadas pelas características dos institutos produtores e dos dados brutos (insumos), pelos processos de produção (métodos e procedimentos) e pelas práticas de integração de dados quando os produtos são agregados de fontes diferentes (*multisource*).

Para Aracri *et al*. (2018), a heterogeneidade das fontes de *big data* e o fato de que suas estruturas dependem da aplicação às quais servem – e não para fins estatísticos – também impõem desafios para o processo de checagem da qualidade. Além disso, questionam a possibilidade de especificação dos requisitos de qualidade em contextos de obscuro entendimento sobre qual semântica os dados devem trazer. Tal reflexão faz lembrar que além das características dos insumos e processos pelos quais os dados passaram, o propósito de sua utilização e os interesses do usuário também devem ser considerados quando da aferição da qualidade. Loshin (2014) vai além das questões de importância e utilidade e destaca que um mesmo conjunto de dados pode possuir diferentes significados em diferentes contextos, levantando também questões de validade e consistência. Ou seja, aparentemente, diretrizes únicas e regras gerais de decisão não serão suficientes para se trabalhar com a aferição da qualidade estatística em diferentes contextos de *big data* e múltiplas fontes (De Waal, Delden e Scholtus, 2017; Máslankowski e Nowicka, 2018).

Apesar do cenário desafiador, De Waal, Delden e Scholtus (2017) propõem maneiras de mensuração da qualidade em termos de (i) acurácia e (iv) coerência em seu *framework* para avaliação da qualidade de produtos estatísticos derivados de múltiplas fontes – considerando exclusivamente pesquisas do tipo *survey* e registros administrativos. Para lidar com a heterogeneidade das fontes, consideram seis diferentes configurações de bases de dados a partir dos aspectos: nível de agregação; unidades de pesquisa; variáveis mensuradas; cobertura; aspectos temporais; população-alvo; e tipo de dado.

A título de conhecimento adicional, externo ao escopo das estatísticas públicas e oficiais, Fagundes, Macedo e Freund (2017) e Jaya *et al*. (2017) realizam revisões de literatura sobre o processo de transição da qualidade de dados para a de *big data* sob a ótica da ciência da informação. No âmbito da computação, Gudivada, Apon e Ding (2017) fazem uma importante discussão sobre as dimensões de qualidade de dados para *big data* e *Machine Learning*. Na tecnologia da informação, Aracri *et al*. (2018) defendem uma abordagem baseada em ontologia para aferição da qualidade de dados.

4.3 *Frameworks* de qualidade no contexto do *big data*

Os *frameworks* tradicionais de qualidade utilizados no contexto de estatísticas públicas – e, consequentemente, indicadores socioeconômicos – consideram predominantemente dados derivados de pesquisas amostrais do tipo *survey*. No âmbito da utilização de dados do tipo *big data*, os *frameworks* construídos para contextos de dados oriundos de registros administrativos correspondem àqueles que parecem mais se adequar, uma vez que fontes administrativas e de *big data* compartilham uma série de características (Reis *et al*., 2016).

Reinert e Stoltze (2016) fazem uma extensa revisão – com testes práticos – em treze proposições de métodos – orientações, *checklists*, *frameworks* – para aferição da qualidade de registros administrativos, construídos entre 2009 e 2014 por instituições de diversos países e uniões continentais. Conforme os autores, todas as propostas possuem indicadores agrupados em dimensões e hiperdimensões de qualidade, com algumas apresentando algum outro tipo de categoria. Por mais que diversas propostas possuam dimensões e hiperdimensões comuns, a maioria possui indicadores próprios, provavelmente pelo fato de que foram pensadas para contextos e estatísticas de interesse particular (Brancato *et al*., 2017). A publicação faz parte de um conjunto de entregas do *ESSnet Quality of Multisource Statistics* (Komuso), que, conforme Brancato e Di Consiglio (2018), trata-se de um projeto de pesquisa que está reunindo os principais modelos de aferição da qualidade de registros administrativos para compor um *framework* que relaciona as dimensões de qualidade dos *outputs* de processos estatísticos com as principais fontes de erro em contextos *multisource* (Brancato e Ascari, 2018).

Entretanto, conforme Reis *et al*. (2016), registros administrativos são constituídos em ambientes controlados, em contraposição ao processo de geração nem sempre controlado de *big data*, tornando necessária a constituição de *frameworks* específicos para indicadores gerados a partir de *big data*.

Nesse sentido, esta revisão bibliográfica permitiu perceber que alguns *frameworks* já foram propostos para o contexto de *big data*, como: o *Unece Big Data Quality Framework* (Unece, 2015); o *Eurostat Accreditation Procedure for Statistical Data from Non-Official Sources* (Eurostat, 2014a); o *AAPOR Big Data Total Error* (Kreuter *et al*., 2015); o *Conceptual Framework for Quality in Big Data* (Batini *et al*., 2015); e o *Big Data Quality Assessment Framework* (Cai e Zhu, 2015). O quadro 4 faz um resumo comparativo das dimensões de qualidade contempladas em cada proposta, sendo as duas primeiras propostas diretamente associadas à produção de estatísticas públicas e as outras três, mais genéricas.

QUADRO 4
Comparativo entre *frameworks* selecionados de qualidade para estatísticas oficiais e públicas produzidas a partir de *big data*

Framework	Hiperdimensões	Domínios	Dimensões – elementos
BDQF[1] (Unece, 2015)	a. Fonte b. Metadados c. Dados	i. Entrada (origem) ii. Meio (processo) iii. Saída (produto)	Ambiente institucional (i.a.; iii.a.)
			Privacidade e segurança (i.a.; iii.a.)
			Complexidade (i.b.; iii.b.)
			Completitude (i.b.)
			Usabilidade (i.b.)
			Coerência e consistência (i.b.; i.c.; iii.c.); e linkabilidade (i.b.; ii.c.; iii.c.)
			Validade (i.b; i.c; iii.c.)
			Aspectos temporais (i.b.; iii.c.)
			Acurácia e seletividade (i.c.; iii.c.)
			Acessibilidade e clareza (iii.b.)
			Relevância (iii.b.)
			Ambiente institucional (i.a.; iii.a.)
APNOS[2] (Eurostat, 2014a)	a. Fonte (conteúdo) b. Metadados c. Dados agregados d. Microdados e. Fonte (instituição)	Entrada (origem)	Uso esperado
			Usabilidade
			Cooperação
		Saída (produto)	Relevância
			Acurácia e confiabilidade
			Pontualidade e oportunismo temporal
			Coerência e comparabilidade
			Acessibilidade e clareza
			Novidade

(Continua)

(Continuação)

Framework	Hiperdimensões	Domínios	Dimensões – elementos
BDTE[3] (Kreuter *et al*., 2015)	Não explícitas	Geração (origem)	Sinal com ruído
			Sinal perdido
			Incompletude
			Seleção não aleatória
			Metadados faltantes
		ETL (extração, transformação, carregamento)	Especificação
			Combinação
			Codificação
			Edição
			Data munging
			Integração de dados
		Análise	Acúmulo de ruídos
			Correlações espúrias
			Endogeneidade incidental
CFQBD[4] (Batini *et al*., 2015)	Não explícitas	Não explícitos	Acurácia – acurácia, correção, validade e precisão
			Completude – completude, pertinência e relevância
			Redundância – redundância, minimalidade, compactude e concisão
			Legibilidade – legibilidade, compreensibilidade, clareza e simplicidade
			Acessibilidade – acessibilidade e disponibilidade
			Consistência – consistência, coesão e coerência
			Veridicidade – veridicidade, credibilidade, confiabilidade e reputação
BDQAF[5] (Cai e Zhu, 2015)	Não explícitas	Não explícitos	Disponibilidade – acessibilidade, temporalidade e autorização
			Usabilidade – definição/documentação, credibilidade e metadados
			Confiabilidade – acurácia, integridade, consistência, completude e auditabilidade
			Relevância – adequação ao uso
			Apresentação – legibilidade e estrutura

Elaboração dos autores.
Notas: [1] BDQF – *Big Data Quality Framework*.
[2] APNOS – *Accreditation Procedure for Statistical Data from Non-Official Sources*.
[3] BDTE – *Big Data Total Error*.
[4] CFQBD – *Conceptual Framework for Quality in Big Data*.
[5] BDQAF – *Big Data Quality Assessment Framework*.

O *framework* proposto por Unece (2015) – BDQF – é reconhecido pelos proponentes como preliminar e de cunho orientador às práticas de aferição da qualidade estatística. É idealizado a partir de uma adaptação dos *frameworks* tradicionais orientados para a qualidade de dados administrativos, considerando os domínios

de *input*, *process* e *output*, com três hiperdimensões – fonte, metadados e dados. Sua estrutura sugere dimensões e indicadores que podem ser utilizados na avaliação da qualidade dos insumos mesmo antes dos dados disponíveis (nas hiperdimensões fonte e metadados), fornecendo a ideia de qualidade potencial dos indicadores e demais produtos estatísticos (Reis *et al.*, 2016). No âmbito do processo, o *framework* não é concreto, trazendo apenas princípios gerais, sem sugestão de indicadores.

A proposta de Eurostat (2014a), intitulada APNOS também é uma extensão de *frameworks* para estatísticas oficiais geradas a partir de registros administrativos, passando a considerar qualquer outro tipo de dado secundário (Reis *et al.*, 2016). Sua estrutura se utiliza de um procedimento de aplicação subdividido em cinco estágios. O primeiro estágio pode ser conduzido sem a posse de dados e fornece uma percepção sobre a potencialidade em termos de cobertura, unidades e variáveis, temporalidade e frequência. O segundo explora as possibilidades de obtenção dos dados, passando por níveis de agregação e formato. No terceiro estágio, a qualidade dos dados é avaliada de forma concreta. O quarto estágio envolve uma análise de custo-benefício e de riscos, orientando a continuidade da utilização do *big data*. Por fim, o último estágio corresponde ao firmamento de acordo de cooperação entre o provedor do *big data* e o produtor interessado.

O *framework* proposto por Kreuter *et al.* (2015), o BDTE, é uma construção da *American Association for Public Opinion Research* (AAPOR), organização dedicada a estudos no campo da opinião pública – atitudes, normas, valores e comportamentos. Corresponde a uma extensão do *Total Survey Error* (TSE), passando a considerar fontes de erros específicas de *big data* (Brancato e Di Consiglio, 2018). Este considera especificamente a acurácia, em termos e viés e erro-padrão, dos produtos gerados, avaliando as bases de dados em termos de possíveis erros presentes nas "linhas, colunas e células" em três etapas: geração dos dados; extração, transformação e carregamento (*extract, transform, load* – ETL); e análise.

Batini *et al.* (2015) propuseram o CFQBD, que mede a qualidade do ponto de vista do usuário e considera, além das dimensões, três coordenadas que orientam e determinam a avaliação da qualidade: o tipo dos dados, a fonte dos dados e o domínio de aplicação. Ou seja, os autores compreendem que os critérios para mensuração dependem do contexto para o qual estão sendo aplicados (Fagundes, Macedo e Freund, 2017). Destacam, por exemplo, que a acurácia depende do tipo do dado, realizando testes em mapas, dados relacionais, *linked open data* e textos não estruturados – percebendo impactos no que concerne à dimensão completude e também a outras dimensões. Citro (2014) corrobora esse entendimento ao afirmar que a mensuração de alguns componentes da acurácia baseados no TSE (não respostas, erro de mensuração, erro de processamento de dados etc.) propostos por Biemer *et al.* (2014) não se aplicam a dados oriundos da interação humana com redes sociais virtuais e mecanismos de buscas na internet, por exemplo.

Assim como o CFQBD, a proposta de Cai e Zhu (2015), o BDQAF também foca na qualidade da perspectiva do usuário e da satisfação dos seus interesses. Os autores consideram cinco dimensões subdivididas em elementos e sugerem alguns indicadores para mensuração de cada aspecto. Além do *framework*, disponibilizaram também um *Quality Assessment Process for Big Data*.

A aplicabilidade e a suficiência dos *frameworks* identificados carecem de ser mais profundamente testadas em situações reais diversas. O que se pode perceber de antemão é que nenhuma proposta cobre todas as dimensões percebidas na subseção 4.2. Além disso, alguns estudos sugerem a necessidade de expansão desses *frameworks*, conforme mostra a subseção 4.4.

Assim, fica como sugestão de ponto de partida para estudos futuros a realização de testes práticos de aplicação dos *frameworks* aqui identificados em diferentes contextos, para que se possa perceber quais aspectos estruturais estão desempenhando satisfatoriamente e quais necessitam ser melhorados ou alterados.

4.4 Lacunas de pesquisa

Alguns trechos retirados da literatura investigada podem diretamente sinalizar lacunas de pesquisa, conforme sintetiza o quadro 5.

QUADRO 5
Trechos associados a lacunas de pesquisa retirados das publicações investigadas

Publicação	Trecho	Página
Barcaroli, Golini e Righi (2018)	"É crucial a definição de um *framework* metodológico que permita o uso eficiente de diferentes fontes, e também a avaliação da acurácia das estimativas obtidas na abordagem *model-based*".	2
Brancato e Ascari (2018)	"Alguns tópicos relativos à confiabilidade, coerência e comparabilidade necessitam de análises e experimentações mais aprofundadas".	8
Brancato e Di Consiglio (2018)	"*Frameworks* teóricos para *multisource* e *big data* devem ser traduzidos em termos operacionais, contendo medidas e indicadores relevantes e informativos".	8
Brancato *et al*. (2018)	"Necessidade de se repensar uma estratégia de qualidade eficiente e direcionada".	4
	"Planejar métricas de mensuração da qualidade no novo ambiente de produção estatística".	9
De Waal *et al*. (2018)	"Um tópico importante de trabalho futuro é o desenvolvimento de um *framework* sistemático para situações, métodos e métricas que podem surgir em contextos de múltiplas fontes" .	10
Hajnovic (2018)	"Por ser um campo emergente, há pouca orientação para a mensuração da qualidade das aplicações de *big data* e *data science* em estatísticas oficiais".	1
Hand (2018)	"Construção de métricas de qualidade e *scorecards* de qualidade para bases de dados".	10
Máslankowski e Nowicka (2018)	"Não há um *framework* unificado para qualidade de *big data* que pode ser aplicado para diferentes tipos de bases de dados".	1
	"A variedade de *frameworks* de qualidade para *big data* a disposição permite criar um conjunto de indicadores capaz de aferir diferentes aspectos da usabilidade de uma base de dados".	1
	"A solução está na criação de diferentes *frameworks* dependentes dos dados em utilização".	1

(Continua)

(Continuação)

Publicação	Trecho	Página
Salgado *et al*. (2018)	"O QAF deve ser revisado em termos de potenciais novos indicadores".	10
	"O paradigma do TSE precisa ser adaptado para o contexto de dados de registros telefônicos".	10
De Waal, Delden e Scholtus (2017)	"Construir métricas e métodos para computá-las mais adequados para aplicações práticas reais".	44
	"Aprofundar testes de adequação de indicadores propostos para a dimensão coerência no contexto de múltiplas fontes".	44
	"Ampliar a quantidade de configurações de dados examinadas no projeto Komuso da ESSnet".	45
	O autor apresenta também uma tabela com outros *gaps* pontuais e específicos.	46
Nasem (2017b)	"Novas fontes de dados requerem expansão e desenvolvimento adicional dos *frameworks* de qualidade existentes para que novos componentes e aspectos possam ser incluídos e enfatizados".	126
Hackl (2016)	"Um *framework* e critérios de qualidade necessitam ser desenhados ou adaptados para os vários tipos de *big data* e de produtos estatísticos que são baseados nesses dados".	51
Reinert e Stoltze (2016)	"Necessidade de métodos padronizados que meçam a qualidade do *input* e que contenham indicadores quantitativos claros, especialmente para a dimensão acurácia".	36
	"Está ficando claro que uma única lista de indicadores para todos os diferentes contextos de aplicação não se faz viável. Temos nos concentrado em conjuntos simples de indicadores bem adequados para o monitoramento contínuo da qualidade no contexto de uma fonte específica".	36
	"Necessidades de trabalhos futuros que definam possíveis erros de cobertura".	36
	"Trabalhos futuros podem focar no desenvolvimento de uma lista de indicadores que auxilie na avaliação da qualidade geral de uma fonte de dados antes mesmo da obtenção dos dados".	37
Reis *et al*. (2016)	"Necessidade de complementar as diferentes propostas (BDQF, APNOS e BDTE) e de estruturar os *links* entre a qualidade do *input* e do *process* com a qualidade do *output*".	16
	"Elementos de cada um dos *frameworks* analisados podem ser combinados para gerar um *framework* de maior qualidade".	16
	"Novas dimensões de qualidade, como complexidade, são também importantes, assim como a proposição de novos indicadores de qualidade".	16
Agafitei *et al*. (2015)	"O ESS QAF ainda não é suficiente para cobrir a complexidade do processo de produção que se utiliza de múltiplas fontes de dados e métodos".	207

Elaboração dos autores.

Os trechos são atuais e sugerem alterações ou expansões nos *frameworks* de qualidade para produtos estatísticos gerados a partir de *big data* até então propostos. Em geral, apontam para a necessidade de incorporação de novas dimensões de qualidade, de desenvolvimento de métricas viáveis e de flexibilização dos *frameworks* para bom funcionamento em contextos com diferentes características.

5 CONSIDERAÇÕES FINAIS

Neste estudo, pôde-se perceber a relevância do tema por sua atualidade e pela crescente produção técnico-científica, associada predominantemente a órgãos europeus de estatística e de pesquisa vinculados ao setor público e ao terceiro setor. Tal constatação se apresenta alinhada com os pressupostos iniciais de que a Revolução

dos Dados vem impactando na forma como o conhecimento é produzido e também de que os produtores de estatísticas públicas e oficiais são os principais interessados nesse processo de inovação na construção de indicadores socioeconômicos. No Brasil, entretanto, estudos na área são ainda incipientes.

No que diz respeito à aferição da qualidade de indicadores produzidos a partir de dados digitais, ou *big data*, os estudos analisados, em geral, têm considerado a necessidade de se observar aspectos relativos a diversas dimensões em três domínios – *input*, *process* e *output*. Alguns estudos, entretanto, propõem uma aferição focada nas dimensões do *output* por questões de viabilidade. Tal proposição sustenta-se na justificativa de que dimensões de *input* e *process* são difíceis de se medir, especialmente em contextos *multisource*, muito comuns quando se trabalha com dados oriundos de *big data*.

Para tanto, alguns *frameworks* já foram propostos dentro e fora da realidade da produção de estatísticas públicas: o *UNECE Big Data Quality Framework*; o *Eurostat Accreditation Procedure for Statistical Data from Non-Official Sources*; o *AAPOR Big Data Total Error*; o *Conceptual Framework for Quality in Big Data*; e o *Big Data Quality Assessment Framework.*

Entretanto, pôde-se perceber que cada contexto de construção de um indicador socioeconômico a partir de *big data* pode possuir características diferentes – diferenças relativas predominantemente ao tipo e às particularidades de insumos (dados brutos) e instituições fornecedoras, aos processos de produção estatística e aos processos de integração de dados de múltiplas fontes. Dessa forma, definir um conjunto único e estático de dimensões em um *framework* para aferição da qualidade de produtos estatísticos pode não funcionar, uma vez que a adequabilidade das dimensões a cada contexto pode variar.

Em adição, a qualidade de um produto estatístico depende de a quem tal produto interessa. Usuários e produtores estatísticos geralmente observam a qualidade de maneiras diferentes, de acordo com seus requisitos e propósitos de utilização. Além disso, a um mesmo dado pode se atribuir semântica diferenciada perante distintos vieses. Esses fatores acarretam, novamente, uma variação nas dimensões de qualidade que se fazem adequadas e relevantes para cada contexto, corroborando a tese de que *frameworks* únicos e estáticos não são suficientes.

Nesse contexto, a quase totalidade das mais recentes lacunas de pesquisa identificadas neste estudo defendem que os *frameworks* até então apresentados não são definitivos, necessitando de ajustes ou expansões que viabilizem uma cobertura maior e mais clara das dimensões de qualidade – capturando novos aspectos e de forma mais clara, especialmente nas tradicionais dimensões acurácia e confiabilidade e coerência e comparabilidade, com métricas e indicadores objetivos e viáveis de serem mensurados – assim como uma adaptabilidade que permita atender às especificidades de cada diferente cenário de aplicação.

REFERÊNCIAS

ABNT – ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR ISO 9000:2005**: sistemas de gestão da qualidade – fundamentos e vocabulário. Rio de Janeiro: ABNT, 2005.

ABS – AUSTRALIAN BUREAU OF STATISTICS. **The ABS data quality framework**. [s.l.]: ABS, 2009.

AGAFITEI, M. *et al*. Measuring output quality for multisource statistics in official statistics: some directions. **Statistical Journal of the IAOS**, v. 31, n. 2, p. 203-211, 2015.

AL-HAJJAR, D. *et al*. Framework for social media big data quality analysis. *In*: BASSILIADES, N. *et al*. (Ed.). **New trends in database and information systems II**. Cham: Springer, 2015. v. 312. p. 301-314.

ARACRI, R. M. *et al*. On the experimental usage of ontology-based data management for the Italian integrated system of statistical registers: quality issues. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

ASQUER, A. **The governance of big data**: perspectives and issues. *In*: INTERNATIONAL CONFERENCE ON PUBLIC POLICY, 1., 2013, Grenoble. **Proceedings**… Grenoble: IPPA, 2013.

BAKER, R. Big data: a survey research perspective. *In*: BIEMER, P. *et al*. (Ed.) **Total survey error in practice**. Hoboken: Wiley, 2017. p. 47-70.

BARCAROLI, G.; GOLINI, N.; RIGHI, P. Quality evaluation of experimental statistics produced by making use of big data. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

BATINI, C. *et al*. Methodologies for data quality assessment and improvement. **ACM Computing Surveys**, v. 41, n. 3, art. 16, p. 1-52, 2009.

BATINI, C. *et al*. From data quality to big data quality. **Journal of Database Management**, v. 26, n. 1, p. 60-82, 2015.

BERGAMASCHI, S. *et al*. Big data research in Italy: a perspective. **Engineering**, v. 2, n. 2, p. 163-170, 2016.

BIEMER, P. P. Total survey error: design, implementation, and evaluation. **The Public Opinion Quarterly**, v. 74, n. 5, p. 817-848, 2010.

BIEMER, P. *et al*. A system for managing the quality of official statistics, with discussion. **Journal of Official Statistics**, v. 30, n. 3, p. 381-442, 2014.

BIEMER, P. *et al*. (Ed.) **Total survey error in practice**. Hoboken: Wiley, 2017.

BIFFIGNANDI, S.; SIGNORELLI, S. From big data to information: statistical issues through examples. *In*: SCIENTIFIC MEETING OF THE CLASSIFICATION AND DATA ANALYSIS GROUP, 10., 2015, Cagliari. **Proceedings**... Cagliari: SIS, 2015. Disponível em: <https://bit.ly/3yLcZeG>.

BOSSEL, H. **Indicators for sustainable development**: theory, method, applications. Winnipeg: IISD, 1999.

BOSSOI, R. A. C. A proteção dos dados pessoais face às novas tecnologias. *In*: ENCONTRO NACIONAL CONPEDI/UFSC, 23., 2014, Florianópolis. **Anais**... Florianópolis: Conpedi, 2014.

BRAAKSMA, B.; ZEELENBERG, K. Re-make/re-model: should big data change the modelling paradigm in official statistics? **Statistical Journal of the IAOS,** v. 31, n. 2, p. 193-202, 2015.

BRACKSTONE, G. Managing data quality in a statistical agency. **Survey Methodology**, v. 25, n. 2, p. 1-23, 1999.

BRANCATO, G.; ASCARI, G. Guidelines on the quality of multisource statistics. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

BRANCATO, G. *et al*. **Guidelines on the quality of multisource statistics – Outline of the quality guidelines document**. [s.l.]: Eurostat, 2017. Disponível em: <https://bit.ly/2SwhSr9>. Acesso em: 28 jan. 2019.

BRANCATO, G. *et al*. The new quality strategy in the modernised Italian National Statistical Institute. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

BRANCATO, G.; DI CONSIGLIO, L. Conceptualising quality for big data. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

BRYMAN, A. **Social research methods**. 2nd ed. Oxford: Oxford University Press, 2004.

BUELENS, B. *et al*. **Shifting paradigms in official statistics**: from design-based to model-based to algorithmic inference. The Hague; Heerlen: Statistics Netherlands, 2012. (Discussion Paper, 201218).

CAI, L.; ZHU, Y. The challenges of data quality and data quality assessment in the big data era. **Data Science Journal**, v. 14, n. 2, p. 1-10, 2015.

CARVALHO, M. A. ***Framework* conceitual para ambiente virtual colaborativo das comunidades virtuais de prática nas universidades no contexto de e-gov**. 2013. Tese (Doutorado) – Universidade Federal de Santa Catarina, Florianópolis, 2013.

CERRONI, F.; DI BELLA, G.; GALIÈ, L. Evaluating administrative data quality as input of the statistical production process. **Rivista di Statistica Ufficiale**, v. 16, n. 1-2, p. 117-146, 2014.

CERVERA, J. L. *et al*. **ESS big data event – Technical Workshop Report**. Rome: Eurostat, 2014.

CITRO, C. F. From multiple modes for surveys to multiple data sources for estimates. **Survey Methodology**, v. 40, n. 2, p. 137-161, 2014.

COCHRAN, W. G. **Sampling techniques**. 3rd ed. New York: Wiley, 1977.

COUPER, M. P. Is the sky falling? New technology, changing media, and the future of surveys. **Survey Research Methods**, v. 7, n. 3, p. 145-156, 2013.

CRESWELL, J. W. **Projeto de pesquisa**: métodos qualitativo, quantitativo e misto. 3. ed. Porto Alegre: Artmed, 2010.

CROSBY, P. B. **Quality is free**: the art of making quality certain. New York: McGraw-Hill, 1979.

DAAS, P. J. H. Big data and official statistics: sharing advisory board. **Software Sharing Newsletter**, n. 7, p. 2-3, 2012.

DAAS, P. J. H. *et al*. **Checklist for the quality evaluation of administrative data sources**. The Hague; Heerlen: Statistics Netherlands, 2009. (Discussion Paper, n. 09042).

DAAS, P. J. H. *et al*. Data science and the future of statistics. *In*: DATA SCIENCE NL MEETUP, 1., 2012, Utrecht. **Proceedings**… Utrecht: Data Science NL, 2012.

DAAS, P. J. H. *et al*. Big data and official statistics. *In*: NEW TECHNIQUES AND TECHNOLOGIES IN STATISTICS CONFERENCE, 2013, Brussels. **Proceedings**… Brussels: Eurostat, 2013.

DAAS, P. J. H. *et al*. Big data as a source for official statistics. **Journal of Official Statistics**, v. 31, n. 2, p. 249-262, 2015.

DAAS, P. J. H.; PUTS, M. J. H. Big data as a source of statistical information. **The Survey Statistician**, n. 69, p. 22-31, 2014.

DAVENPORT, T. H.; PATIL, D. Data scientist: the sexiest job of the 21st century. **Harvard Business Review**, v. 90, n. 10, p. 70-77, 2012.

DE JONGE, E.; VAN PELT, M.; ROOS, M. **Time patterns, geospatial clustering and mobility statistics based on mobile phone network data**. The Hague; Heerlen: Statistics Netherlands, 2012. (Discussion Paper, n. 201214).

DEMUNTER, C. Tourism statistics: early adopters of big data? *In*: UNWTO INTERNATIONAL CONFERENCE ON TOURISM STATISTICS, 6., 2017, Manila. **Proceedings**… Manila: UNWTO, 2017.

DE WAAL, T.; DELDEN, A. van; SCHOLTUS, S. **Work package 3**: framework for the quality evaluation of statistical output based on multiple sources. [s.l.]: Eurostat, 2017.

_____. Quality measures and indicators for multisource statistics. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018. Disponível em: <https://bit.ly/3bPJfDE>. Acesso em: 21 maio 2021.

DUNNE, J. Big data coming soon... to an NSI near you. *In*: ISI WORLD STATISTICS CONGRESS, 59., 2013, Hong Kong. **Proceedings**… Hong Kong: ISI, 2013.

ECB – EUROPEAN CENTRAL BANK. **ECB Statistics Quality Framework (SQF)**. [s.l.]: ECB, Apr. 2008. Disponível em: <https://bit.ly/3xZaTHv>. Acesso em: 28 jan. 2019.

ERCOLE, F. F.; MELO, L. S. de; ALCOFORADO, C. L. G. C. Revisão integrativa *versus* revisão sistemática. **Revista Mineira de Enfermagem**, v. 18, n. 1, p. 9-12, 2014.

ESSNET. **Use of administrative and accounts data in business statistics – Deliverable 6.3**: guidance on the accuracy of mixed-source statistics. [s.l.]: ESSnet, 2013. (Working Papers, n. 6). Disponível em: <https://bit.ly/3y4BcMr>. Acesso em: 28 jan. 2019.

_____. **Some quality aspects and future prospects for the production of official statistics with mobile phone data**. [s.l.]: ESSnet, 2018. Disponível em: <https://bit.ly/2QLMVyK>. Acesso em: 28 jan. 2019.

EUROSTAT. **Scheveningen memorandum on big data and official statistics**. [s.l.]: Eurostat, 2013. Disponível em: <https://bit.ly/3fhg3aI>. Acesso em: 28 jan. 2019.

_____. **Accreditation procedure for statistical data from non-official sources**. [s.l.]: Eurostat, 2014a. Disponível em: <https://bit.ly/3vVL7SF>. Acesso em: 28 jan. 2019.

_____. Big data: an opportunity or a threat to official statistics? *In*: CONFERENCE OF EUROPEAN STATISTICIANS, 62., 2014, Paris. **Proceedings**… Paris: Eurostat, 2014b.

_____. **Quality Assurance Framework of the European Statistical System**. [s.l.]: ESS, 2015. Disponível em: <https://bit.ly/3tBIFzk>. Acesso em: 28 jan. 2019.

_____. **Report from the commission to the European parliament and the council on the implementation of the European Statistics Code of Practice and coordination within the European Statistical System**. [s.l.]: Eurostat; Comissão Europeia, 2016. Disponível em: <https://bit.ly/3hQiXok>. Acesso em: 21 maio 2021.

_____. **European statistics code of practices**: for the National Statistical Authorities and Eurostat (EU statistical authority). Luxembourg: Eurostat, 2017. Disponível em: <https://bit.ly/3hghnMh>. Acesso em: 28 jan. 2019.

FAGUNDES, P. B.; MACEDO, D. D. J. de; FREUND, G. P. A produção científica sobre qualidade de dados em *big data*: um estudo na base de dados Web of Science. **RDBCI – Revista Digital de Biblioteconomia e Ciência da Informação**, v. 16, n. 1, p. 194-210, 2017.

FAO – FOOD AND AGRICULTURE ORGANIZATION. **The FAO Statistics Quality Assurance Framework:** implementation strategy and plan (draft). Rome: FAO, 2014.

FLEKOVA, L.; GUREVYCH, I. Can we hide in the web? Large scale simultaneous age and gender author profiling in social media. *In*: CONFERENCE AND LABS EVALUATION FORUM, 2013, Valencia. **Proceedings**… Valencia: Promise, 2013.

FLORESCU, D. *et al*. **Will 'big data' transform official statistics?** *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 2014, Vienna. **Proceedings**… Vienna: Statistics Austria; Eurostat, 2014. Disponível em: <https://bit.ly/3o70hlC>. Acesso em: 28 jan. 2019.

FMI – FUNDO MONETÁRIO INTERNACIONAL. The Data Quality Assessment Framework. **FMI**, June 25, 2003. Disponível em: <https://bit.ly/3uHzZc8>. Acesso em: 28 jan. 2019.

FORBES, D. A. Strategies for managing behavioural symptomatology associated with dementia of the Alzheimer type: a systematic overview. **Canadian Journal of Nursing Research**, v. 30, n. 2, p. 67-86, 1998.

FRY, B. **Visualizing data**: exploring and explaining data with the processing environment. Sebastopol: O'Reilly Media, 2008.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GREEN, S.; HIGGINS J. **Cochrane handbook for systematic reviews of interventions 4.2.5**. [s.l.]: The Cochrane Collaboration, 2005.

GROVES, R.; LYBERG, L. Total survey error: past, present and future. **Public Opinion Quarterly**, v. 74, n. 5, p. 849-879, 2010.

GUDIVADA, V.; APON, A.; DING, J. Data quality considerations for big data and machine learning: going beyond data cleaning and transformations. **International Journal on Advances in Software**, v. 10, n. 1, p. 1-20, 2017.

GURWITZ, P. Combining data mines and attitude research. *In*: KADEN, R. **Leading edge marketing research**: 21st-century tools and practices. Washington: Sage, 2012. p. 50-70.

HACKL, P. Big data: what can official statistics expect? **Statistical Journal of the IAOS**, v. 32, n. 1, p. 43-52, 2016.

HAJNOVIC, F. Measuring the quality of commercial and big data sources for official statistics. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

HAND, D. J. Official statistics in the new data ecosystem. *In*: NEW TECHNIQUES AND TECHNOLOGIES IN STATISTICS CONFERENCE, 2015, Brussels. **Proceedings**… Brussels: Eurostat, Mar. 2015.

_____. Statistical challenges of administrative and transaction data (with discussion). **Journal of the Royal Statistical Society**, v. 181, n. 3, p. 555-605, 2018.

HARWOOD, A.; MAYER, A. Big data and semantic technology: A future for data integration, exploration and visualisation. **Statistical Journal of the IAOS**, v. 32, n. 4, p. 613-626, 2016.

HASSANI, H.; SAPORTA, G.; SILVA, E. S. Data mining and official statistics: the past, the present and the future. **Big Data**, v. 2, n. 1, p. 34-43, 2014.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. 2nd ed. New York: Springer, 2009.

HELBING, D. *et al*. **Behavioural control or digital democracy?** It's time to decide. [s.l.]: ResearchGate, 2016.

HILBERT, M. Big data for development: a review of promises and challenges. **Development Policy Review**, v. 34, n. 1, p. 135-174, 2016.

HIQA – HEALTH INFORMATION AND QUALITY AUTHORITY. **Background paper to support guidance for a data quality framework for he-**

alth and social care. Mahon; Cork: HIQA, 2018. Disponível em: <https://bit.ly/3eBGDuU>. Acesso em: 28 jan. 2019.

HSIEH, Y. P.; MURPHY, J. Total Twitter error: decomposing public opinion measurement on Twitter from a total survey error perspective. *In*: BIEMER, P. *et al*. (Ed.) **Total survey error in practice**. Hoboken: Wiley, 2017. p. 23-46.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Código de boas práticas das estatísticas do IBGE**. Rio de Janeiro: IBGE, 2013.

IWIG, W. *et al*. **Data quality assessment took for administrative data**. [s.l.]: U.S. Bureau of Labor Statistics, 2013. Disponível em: <https://bit.ly/2SsjaUk>. Acesso em: 28 jan. 2019.

JANNUZZI, P. de M. **Indicadores sociais no Brasil**: conceitos, fontes de dados e aplicações. Campinas: Editora Alínea, 2001.

_____. Considerações sobre o uso, mau uso e abuso dos indicadores sociais na formulação e avaliação de políticas públicas municipais. **Revista de Administração Pública**, v. 36, n. 1, p. 51-72, 2002.

JANNUZZI, P. de M.; CRACIOSO, L. de S. Produção e disseminação da informação estatística pelas agências estaduais no Brasil. **Revista São Paulo em Perspectiva**, v. 16, n. 3, p. 92-103, 2002.

JAYA, I. M. *et al*. A review of data quality research in achieving high data quality within organization. **Journal of Theoretical and Applied Information Technology**, v. 95, n. 12, p. 2647-2657, 2017.

JURAN, J.; GRYNA, F. **Quality planning and analysis**. 2nd ed. New York: McGrawHill, 1980.

KITCHIN, R. **The data revolution**: big data, open data, data infrastructures and their consequences. London: Sage, 2014.

_____. The opportunities, challenges and risks of big data for official statistics. **Statistical Journal of the IAOS**, v. 31, n. 3, p. 471-481, 2015.

KRÄTKE, F.; BYIERS, B. **The political economy of official statistics**: implications for the data revolution in sub-Saharan Africa. Maastricht: ECDPM, 2014. (Discussion Papers, n. 170).

KREUTER, F. *et al*. **AAPOR Report on big data**. [s.l.]: Mathematica Policy Research, 2015.

LAURO, B.; TRAVERSO, R. Data fitness for integration. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

LAVALLE, S. Big data, analytics and the path from insights to value. **MIT Sloan Management Review**, v. 52, n. 2, p. 21-32, 2011.

LIU, Z.; JIANG, B.; HEER, J. *imMens*: real-time visual querying of big data. **Computer Graphics Forum**, v. 32, n. 3, p. 421-430, 2013.

LOSHIN, D. **Understanding big data quality for maximum information usability**. [s.l.]: SAS, 2014. Disponível em: <https://bit.ly/3bkPn6i>. Acesso em: 28 jan. 2019.

LUGOMER, K. *et al*. Understanding sources of measurement error in Wi-Fi sensor data in the Smart City. *In*: GIS RESEARCH UK CONFERENCE – GISRUK, 25., 2017, Manchester. **Proceedings**... Manchester: GISRUK, 2017. Disponível em: <https://bit.ly/3eB9oI1>. Acesso em: 28 jan. 2019.

MACFEELY, S. The continuing evolution of official statistics: some challenges and opportunities. **Journal of Official Statistics**, v. 32, n. 4, p. 789-810, 2016.

MANSKI, C. F. **Communicating uncertainty in official economic statistics**. Cambridge, Massachusetts: NBER, 2014. (Working Paper, n. 20098).

MÁSLANKOWSKI, J.; NOWICKA, A. Big data quality issues regarding multidomain statistical data combining: a survey and case studies. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

MAYER-SCHÖNBERGER, V.; CUKIER, K. **Big data**: a revolution that will transform how we live, work, and think. New York: Houghton Mifflin Harcourt, 2013.

MILLER, H. J. The data avalanche is here: shouldn't we be digging? **Journal of Regional Science**, v. 50, n. 1, p. 181-201, 2010.

MILLS, S. *et al*. **Demystifying big data**: a practical guide to transforming the business of government. [s.l.]: TechAmerica Foundation, 2012.

MOHER, D. *et al*. Reprint – Preferred reporting items for systematic review and meta-analyses: the PRISMA Statement. **Physical Therapy**, v. 89, n. 9, p. 873-880, 2009.

NAS – NATIONAL ACADEMY OF SCIENCES. **Frontiers in massive data analysis**. Washington: The National Academies Press, 2013.

NASEM – NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE. **Innovations in federal statistics**: combining data sources while protecting privacy. Washington: The National Academies Press, 2017a.

_____. **Federal statistics, multiple data sources, and privacy protection**: next steps. Washington: The National Academies Press, 2017b.

NEDERPELT, P. **A new model for quality management**. The Hague; Heerlen: Statistics Netherlands, 2010.

NEDERPELT, P.; DAAS, P. **49 factors that influence the quality of secondary data sources**. The Hague; Heerlen: Statistics Netherlands, 2012.

O'CONNOR, B. *et al*. A. From tweets to polls: linking text sentiment to public opinion time series. *In*: INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 4., 2010, Washington. **Proceedings**... Washington: AAAI, 2010.

OCDE – ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Rumo a um ambiente sustentável**: indicadores ambientais. Salvador: OCDE, 2002. (Série Cadernos de Referência Ambiental, v. 9).

_____. **Recommendation of the OECD Council on good statistical practice**. Paris: OCDE, 2015. Disponível em: <https://bit.ly/3hl7ZHc>. Acesso em 28 jan. 2019.

ONS – OFFICE FOR NATIONAL STATISTICS. **Guidelines for measuring statistical output quality**. London: ONS, 2013. Disponível em: <https://bit.ly/2QLcJei>. Acesso em: 28 jan. 2019.

OPHER, A. *et al*. **The rise of the data economy**: driving value through internet of things data monetization. North Castle: IBM, 2016.

PARISE, S.; IYER, B.; VESSET, D. Four strategies to capture and create value from big data. **Ivey Business Journal**, v. 76, n. 4, p. 1-5, 2012.

PLATEK, R.; SARNDAL, C-E. Can a statistician deliver? **Journal of Official Statistics**, v. 17, n. 1, p. 1-20, 2001.

REIMSBACH-KOUNATZE, C. The proliferation of "big data" and implications for official statistics and statistical agencies: a preliminary analysis. Paris: OECD Publishing, 2015. (OECD Digital Economy Papers, n. 245).

REINERT, R.; STOLTZE, P. T. **Checklist for evaluating the quality of input data**. Budapest: Statics Denmark, 2016. Disponível em: <https://bit.ly/3tCYBkA>. Acesso em: 28 jan. 2019.

REIS, F. *et al*. Comparative assessment of three quality frameworks for statistics derived from big data: the cases of Wikipedia page views and automatic identification systems. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 2016, Madrid. **Proceedings**… Madrid: INE; Eurostat, June 2016.

RUDIN, C. *et al*. **Discovery with data**: leveraging statistics with computer science to transform science and society. [s.l.]: ASA, July 2nd, 2014.

SAEBO, H. V. Quality in statistics – from Q2001 to 2016. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 2016, Madrid. **Proceedings**… Madrid: INE; Eurostat, June 2016.

SALGADO, D. *et al*. Estimation of population counts combining official data and aggregated mobile phone data. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

SCANNAPIECO, M.; VIRGILLITO, A.; ZARDETTO, D. Placing big data in official statistics: a big challenge? *In*: NEW TECHNIQUES AND TECHNOLOGIES IN STATISTICS CONFERENCE, 2013, Brussels. **Proceedings**… Brussels: Eurostat, 2013.

SCHMIDT, E.; COHEN, J. **A nova era digital**: como será o futuro das pessoas, das nações e dos negócios. Rio de Janeiro: Intrínseca, 2013.

SCHNORR-BAECKER, S. Statistical monitoring systems to inform policy decision-making, and new data sources. **Statistical Journal of the IAOS**, v. 33, n. 2, p. 1-12, 2016.

SCHUTT, R.; O'NEIL, C. **Doing data science**: straight talk from the frontline. Sebastopol: O'Reilly Media, 2013.

SCHWAB, K. **A quarta revolução industrial**. Tradução de Daniel Moreira Miranda. São Paulo: Edipro, 2016.

SCHWARTZMAN, S. Legitimidade, controvérsias e traduções em estatísticas públicas. **Teoria e Sociedade**, Belo Horizonte, v. 2, p. 9-38, 1997.

SILVA, P. L. N. Statistical thinking and methodology: pillars for quality in the big data era. *In*: IFC CONFERENCE ON STATISTICAL IMPLICATIONS OF THE NEW FINANCIAL LANDSCAPE, 8., 2016, Basel. **Proceedings**… Basel: IFC, 2016.

STATISTICS CANADA. **Statistics Canada quality guidelines**. 5th ed. Ottawa: Statistics Canada, 2009.

STRUIJS, P.; BRAAKSMA, B.; DAAS, P. Official statistics and big data. **Big data and Society**, p. 1-6, Apr.-June 2014.

STRUIJS, P.; DAAS, P. J. H. Big data, big impact? *In*: SEMINAR ON STATISTICAL DATA COLLECTION, 2013, Geneva. **Proceedings**… Geneva: Unece, 2013.

_____. Quality approaches to big data in official statistics. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 2014, Vienna. **Proceedings**… Vienna: Statistics Austria; Eurostat, 2014.

STRUIJS, P.; DE BROE, S. Big data strategies for official statistics. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

TAM, S.-M.; CLARKE, F. Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *In*: INTERNATIONAL CONFERENCE ON BIG DATA FOR OFFICIAL STATISTICS, 2014, Beijing. **Proceedings**… Beijing: UNSD, 2014.

_____. Big data, official statistics and some initiatives by the Australian Bureau of Statistics. **International Statistical Review**, v. 83, n. 3, p. 436-448, 2015a.

_____. **Big data, statistical inference and official statistics**. Canberra: Australian Bureau of Statistics, 2015b. (Research Paper, n. 1351.0.55.054).

TAM, S-M.; KIM, J.-K. Big data ethics and selection-bias: an official statistician's perspective. **Statistical Journal of the IAOS**, v. 34, n. 4, p. 577-588, 2018.

TAO, C.; GAO, J. Quality assurance for big data application: issues, challenges, and needs. *In*: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING, 28., 2016, Redwood City. **Proceedings**… Redwood City: Seke, 2016.

TENNEKES, M.; DE JONGE, E.; DAAS, P. J. H. **Visual profiling of large statistical datasets**. The Hague; Heerlen: Statistics Netherlands, 2011.

THOMPSON, M. E. Dynamic data science and official statistics. **The Canadian Journal of Statistics**, v. 46, n. 1, p. 10-23, 2018.

TRIVINOS, A. N. S. **Introdução à pesquisa em ciências sociais**: a pesquisa qualitativa em educação. São Paulo: Atlas, 1992.

TUFEKCI, Z. Big data: pitfalls, methods and concepts for an emergent field. **SSRN Electronic Journal**, Mar. 7, 2013.

UK STATISTICS AUTHORITY. **Code of practice for statistics**. 2. ed. London: UK Statistics Authority, 2018. Disponível em: <https://bit.ly/3fFZaFr>. Acesso em: 28 jan. 2019.

ULUWIYAH, A. Trusted big data for official statistics: study case – Statistics Indonesia (BPS). *In*: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY SYSTEMS AND INNOVATION, 2016, Bandung. **Proceedings**… Bandung: IEEE, 2016.

UN – UNITED NATIONS. **Guidelines for the template for a Generic National Quality Assurance Framework (NQAF)**. [s.l.]: UN, 2012. Disponível em: <https://bit.ly/3iE449d>. Acesso em: 28 jan. 2019.

_____. **A world that counts**: mobilising the data revolution for sustainable development. New York: UN, 2014.

_____. **UN statistics quality assurance framework**: including a generic statistical quality assurance framework for a UN agency. [s.l.]: UN-SQAF, 2018. Disponível em: <https://bit.ly/3bNJf76>. Acesso em: 28 jan. 2019.

UNECE – UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE. **Fundamental principles of official statistics**. [s.l.]: Unece, 1992. Disponível em: <https://bit.ly/3tApcin>. Acesso em: 28 jan. 2019.

_____. **Generic Statistical Business Process Model (GSBPM)**. [s.l.]: Unece, 2013a. Disponível em: <https://bit.ly/3o6dtXD>. Acesso em: 28 jan. 2019.

_____. **Classification of types of big data**. [s.l.]: Unece, 2013b. Disponível em: <https://bit.ly/3tAqLNh>. Acesso em: 28 jan. 2019.

_____. A suggested framework for national statistical offices for assessing the quality of big data. *In*: NEW TECHNIQUES AND TECHNOLOGIES FOR STATISTICS CONFERENCE – NTTS, 2015, Brussels. **Proceedings**… Brussels: Unece, 2015. Disponível em: <https://bit.ly/3ufrUu8>. Acesso em: 21 maio 2021.

UNGP – UNITED NATIONS GLOBAL PULSE. **Big data for development**: challenges and opportunities. New York: UN Global Pulse, 2012.

UNSC – UNITED NATIONS SECURITY COUNCIL. Big data and modernisation of statistical systems. *In*: STATISTICAL COMMISSION, 45., 2014, New York. **Proceedings**… New York: UNSC, 2014. Disponível em: <https://bit.ly/3hhR4FG>. Acesso em: 28 jan. 2019.

VACCARI, C. **Big data in official statistics**. Thesis (PhD) – University of Camerino, Camerino, 2014.

VAJU, S. C.; MESZAROS, M. T. Administrative data and quality: guidelines towards better quality of administrative data. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

VALE, S. International collaboration to understand the relevance of big data for official statistics. **Statistical Journal of the IAOS**, v. 31, n. 2, p. 159-163, 2015.

VALENTE, N. T. Z.; FUJINO, A. Atributos e dimensões de qualidade da informação nas Ciências Contábeis e na Ciência da Informação: um estudo comparativo. **Perspectivas em Ciência da Informação**, v. 21, n. 2, p. 141-167, 2016.

WALLGREN, A.; WALLGREN, B. **Register-based statistics**: statistical methods for administrative data. 2nd ed. Chichester: Wiley, 2014.

WANG, D. J. *et al*. Measurement error in network data: a re-classification. **Social Networks**, v. 34, n. 4, p. 1-14, 2012.

WANG, R. Y.; STRONG, D. M. Beyond Accuracy: what data quality means to data consumers. **Journal of Management Information System**, v. 12, n. 4, p. 5-34, 1996.

WEF – WORLD ECONOMIC FORUM. **Personal data**: the emergence of a new asset class. Geneva: WEF, 2011.

WIRTHMANN, A.; REIS, F. Ethical implications of using big data for official statistics. *In*: EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS, 9., 2018, Kraków, Poland. **Proceedings**… Kraków: Statistics Poland; Eurostat, 2018.

ZHU, H. *et al*. **Data and information quality research**: its evolution and future. Cambridge, Massachusetts: MIT, 2012. (Working Paper CISL, n. 2012-13).

ZIKOPOULOS, P. C. *et al*. **Understanding big data**: analytics for enterprise class hadoop and streaming data. Ney York: McGraw Hill Enterprises, 2012.

Data da submissão em: 22 mar. 2019.

Primeira decisão editorial em: 14 jun. 2019.

Última versão recebida em: 19 jul. 2019.

Aprovação final em: 6 ago. 2019.