SemeAd 2025
XXVIII Seminários em Administração

10, 11, 12, 13 e 14 de novembro de 2025
ISSN 2177-3866

Uso de Dados Sintéticos em Pesquisas de Consumo - Revisão Sistemática de Literatura e Agenda para Estudos Futuros

BRUNO EDUARDO PEREIRA
UNIVERSIDADE FEDERAL DO PARANÁ (UFPR)

ELDER SEMPREBON
UNIVERSIDADE FEDERAL DO PARANÁ (UFPR)

NAIARA JOHNSSON
FAE CENTRO UNIVERSITÁRIO (FAE)

USO DE DADOS SINTÉTICOS EM PESQUISAS ACADÊMICAS DO COMPORTAMENTO DO CONSUMIDOR E MARKETING: REVISÃO SISTEMÁTICA E AGENDA FUTURA

RESUMO: Este estudo realizou uma revisão sistemática da literatura sobre o uso de dados sintéticos gerados por técnicas de inteligência artificial no campo do comportamento do consumidor e do marketing. Utilizando bases como Scopus, Web of Science e IEEE Xplore, foram identificados 9 estudos nas ciências sociais aplicadas e 2 estudos no marketing/comportamento do consumidor, indicando uma escassez significativa de pesquisas aplicadas nos campos. A análise dos artigos revelou que, embora os dados sintéticos sejam amplamente explorados em áreas como saúde e administração de dados, há lacunas importantes na aplicação em pesquisas comportamentais e de marketing, como a preservação da privacidade, redução de custos e aumento do tamanho amostral, além de sugerir direções futuras como o desenvolvimento de *guidelines* éticos, testes A/B com simulações e integração com escalas psicométricas.

Palavras-chaves: *Dados Sintéticos; Metodologias de Pesquisa; Marketing; Comportamento do Consumidor; Modelos Generativos.*

INTRODUÇÃO

Imagine que uma empresa da indústria farmacêutica deseja testar um novo medicamento, mas ainda não pode testá-lo em humanos. Com tal limitação, ela se utiliza de "manequins inteligentes", criados por meio de dados sintéticos gerados pela inteligência artificial, que reproduzem comportamentos estatísticos de pacientes reais sustentado pelo uso e assim, consegue prever reações, ajustar dosagens e validar o protocolo, sem usar um único paciente real (Kokosi & Harron, 2022).

Os dados sintéticos desempenham um papel fundamental em pesquisas, particularmente em áreas como saúde e ciências sociais, fornecendo um meio de analisar dados sem comprometer a privacidade individual. Essa abordagem inovadora permite que pesquisadores gerem conjuntos de dados que imitam dados do mundo real, minimizando os riscos de divulgação, mas mantendo a utilidade analítica comparável à dos dados reais, permitindo que pesquisadores tirem conclusões válidas sem expor sujeitos reais a riscos (Taub *et al.,* 2016). Apesar do desafios ético em questões como o consentimento para o uso de dados reais na criação de conjuntos de dados sintéticos e potenciais vieses nos dados gerados (Dwivedi, 2024), estes dados sinteticamente produzidos vêm reduzindo cada vez mais a necessidade de acesso a dados pessoais, salvaguardando assim a privacidade individual em áreas de pesquisa sensíveis, como saúde, além de serem mais econômicos em larga escala (Kokosi & Harron, 2022).

Essa criação sintética somente é possível com o uso da inteligência artificial, que cada vez mais aprimora as capacidades de geração de dados sintéticos, expandindo suas aplicações em diversos domínios de pesquisa (Timpone & Yang, 2024). Alguma das principais formas de gerações de dados sintéticos podem ser vistas por meio de técnicas que vão desde Árvores de Decisão (*Decision Tree*) e Regressão Florestal (*Forest Regression*) até o uso de modelos generativos (GANs), todos utilizando técnicas de modelos generativos guiados pelo aprendizado de máquina ou *Machine Learning* (Kokosi & Harron, 2022).

Neste contexto da inteligência artificial, os modelos generativos caracterizam-se em arquiteturas de redes neurais de multicamadas, cuja capacidade reside na inferência da

distribuição de um conjunto de dados empíricos, possibilitando a síntese de novas amostras (D'Amico *et al.,* 2023). Nesse âmbito, as Redes Adversariais Generativas (GANs) instauram um ambiente de simulação, no qual modelos e processos estabelecem uma interação dinâmica para a produção de conjuntos de dados inéditos de eventos, ou em outras palavras, essas GANs criam cenários de simulação onde modelos e processos interagem para criar conjuntos de dados completamente novos de eventos, conhecidos como "Dados Sintéticos" (Le Cun *et al.,* 2015).

Os dados sintéticos também podem ser gerados por meio de árvores de decisão e utilizados para mineração de dados e modelos preditivos, evitando afetar ou mesmo acessar os dados originais, além e criar conjuntos de dados maiores (Eno & Thompson, 2008) ou serem gerados por meio de florestas de regressão, ao qual por meio da amostra do valor de uma variável utiliza-se a floresta para prever outros valores, repetindo para todas as variáveis e capturando as relações aprendidas, modelando até mesmo relações complexas e não lineares entre as variáveis (Rankin *et al*., 2020). Porém o uso das GANs é o formato mais utilizado para esses processos, uma técnica que compreende duas redes neurais distinta, uma geradora e uma discriminadora, cujo treinamento ocorre de forma adversarial. O gerador é responsável pela produção de dados sintéticos, as quais são submetidas ao discriminador ao mesmo tempo que os dados reais, enquanto a função do discriminador consiste na identificação das amostras, discernindo entre dados reais e aqueles gerados artificialmente (Goodfellow *et al.,* 2015).

O objetivo dos dados sintéticos é criar um conjunto de dados que se assemelhe aos dados individuais originais e mantenha o mesmo tamanho de amostra (Kokosi & Harron, 2022). Usado principalmente para acelerar resultados de pesquisas, obter dados fidedignos e completos, testar algoritmos estatísticos, gerar conjunto de dados maiores e preservar a privacidade de pessoas, os dados sintéticos vêm sendo cada vez mais utilizados em no campo da saúde (D'Amico *et al*., 2023) e das finanças (Potluru, 2023), mas ainda encontram carências em estudos nos campos do comportamento do consumidor e do marketing, sejam em pesquisas de mercado ou pesquisas acadêmicas, mesmo com o crível incentivo de uso, devido ao aumento da fiscalização dos dados protegidos pelas Leis de Proteção de Dados (LGPD).

Apesar de estudos sobre o uso da inteligência artificial generativa terem se desenvolvidos recentemente no campo do comportamento do consumidor (Hermann & Puntoni, 2024) e no campo do marketing (Grewal *et al.,* 2024), além do uso consolidado da técnica de *Bootstrapping* (Hayes, 2009) na mediação de pesquisas experimentais, uma técnica de reamostragem que constrói uma distribuição amostral empírica de uma estatística no processo inferencial e que pode ser considerada uma simulação de dados sintéticos baseado em dados reais, os estudos que consideram a elaboração de dados sintéticos por meio de técnicas de aprendizado de máquina nestes campos, ainda é escasso.

Ciente de tal uso deste recurso por áreas em que a sensibilidade dos dados e o uso deles é de suma importância para a vida das pessoas, como saúde e finanças, esse estudo buscou entender como os dados sintéticos têm sido utilizados em pesquisas acadêmicas, especialmente na área de comportamento do consumidor e do marketing, identificando lacunas e oportunidades de pesquisas futuras.

A revisão sistemática seguiu um processo metodológico e sistemático para identificar e avaliar criticamente um conjunto escasso e ainda emergente e de pesquisas, os quais investigam sobre o uso de dados sintéticos em pesquisas acadêmicas no campo do comportamento do consumidor e do marketing, não antes sem entender o seu atual estado no campo das ciências social aplicadas, a partir dos quais foram gerados insights coletivos de conhecimento e sugestões para pesquisas futuras sobre o campo. Os resultados buscaram responder as seguintes questões de pesquisa:

RQ1: Qual é o estado atual da literatura sobre o uso de dados sintéticos em pesquisa acadêmicas no campo das ciências sociais aplicadas?

RQ2: Qual é o estado atual da literatura sobre o uso de dados sintéticos em pesquisa nas áreas do comportamento do consumidor e do marketing?

RQ3: Quais são os contribuições alcançadas e lacunas identificadas em pesquisa no campo do comportamento do consumidor e do marketing?

RQ4: Quais as direções para pesquisas futuras e quais os objetivos práticos no uso de dados sintéticos em pesquisas no campo do comportamento do consumidor e do marketing?

O restante deste artigo está organizado da seguinte forma: a próxima seção descreve o método utilizado para coletar e analisar os dados, complementada por uma visão geral do uso dos dados sintéticos em pesquisas no campo das ciências sociais aplicadas e nas principais linhas de pesquisa sobre comportamento do consumidor e marketing, com uma discussão dos resultados conceituais. Na sequência é apresentada uma seção que aborda os tópicos de desafios, lacunas identificadas e direções para pesquisas futuras, além de propor benefícios do uso dos dados sintéticos no campo do comportamento do consumidor e do marketing. Por fim, são apontadas as limitações do estudo e as conclusões do artigo.

REVISÃO SISTEMÁTICA DE LITERATURA

Seguindo o formato de outros estudos no campo das ciências sociais aplicadas (Ceipek, 2019; Galvagno & Dalli, 2014), foi adotada uma abordagem sistemática para a revisão da literatura, a fim de tornar os resultados confiáveis, verificáveis e reprodutíveis, com o objetivo de identificar o estado da arte no uso de dados sintéticos em pesquisas acadêmicas, com foco no campo das ciências sociais aplicadas, mais especificamente, nas áreas do comportamento do consumidor e no marketing, mostrando tendências, metodologias utilizadas, propondo uma agenda de pesquisa futura e explanando objetivos práticos e benéficos de seu uso em pesquisas acadêmicas.

As fontes foram extraídas com base nos bancos de dados Scopus, Web of Science, ScienceDirect, IEEE Xplore, ACM Digital Library e SciELO. Este procedimento buscou garantir a presença dos artigos mais relevantes publicados em periódicos de ponta e editoras importantes como Emerald, Elsevier, IEEE, Springer, Sage, Taylor and Francis e Wiley. Uma coleção da literatura inicial abordou o uso de dados sintéticos no campo das ciências sociais aplicadas, para identificar a atual conjuntura de seu uso neste campo, aprofundando-se, na sequência, nos estudos do comportamento do consumidor e do marketing. Para a primeira tarefa, utilizou-se os termos "*synthetic data*" AND "s*ocial sciences*" OR "*applied social sciences*", pesquisados no campo de "Tópicos" (Título, Resumo e Palavras-chaves) e considerando os filtros de Área (“Ciências Sociais Aplicadas”), Condição (“Revisado por Pares”), Tipo de documento (“Article” ou “Journal article”) e Área do Conhecimento (“Social Sciences”, “Business”, “Marketing” e “Behavioral Sciences”).

A etapa inicial retornou 10 resultados (2005-2025) e devido ao número pequeno, todos os artigos foram acessados de forma completa. Um dos artigos foi retirado por se tratar de uma nota do editor da revista *Managament Science* sobre o uso de dados no periódico. Dessa forma, 9 foram mantidos por representarem o uso dos dados sintéticos como forma de contribuição em pesquisas e análises estatísticas no campo das ciências sociais. Os resultados apresentados mostram estudos que utilizaram os dados sintéticos para avaliar os riscos de divulgação de identidade em estudos da Covid-19 no Canadá e no estado de Washington nos Estados Unidos

(Eman *et al.,* 2019), contribuindo com a privacidade daqueles que realizavam testes ou estavam contaminados com o vírus.

No mesmo sentido, outro artigo propôs uma nova abordagem para o aumento de dados estatísticos de forma eficientemente, por meio de dados sintéticos gerados pela técnica de LLM (*Large Language Models*) e utilizando dados reais na análise conjunta, validando o resultado por meio de um estudo empírico sobre as preferências de vacina para Covid-19, demonstrando sua capacidade superior de reduzir o erro de estimativa e economizar dados e custos (Wang *et al.,* 2024). Ainda no campo da saúde, um estudo contribuiu com a adaptação de técnicas regionalistas no atendimento de médicos generalistas, por meio da criação de uma Matriz de Dados Sintéticos (MDS), que contribuía no planejamento de atendimento em algumas regiões mais afastadas no Reino Unido, ao qual os pacientes necessitavam de uma atenção primária (Shortt, *et al.,* 2005).

Um estudo sobre atendimentos no campo da saúde investigou o problema de sequenciamento e agendamento de compromissos com tolerâncias de atraso de pacientes, vistos sob incerteza de tempo de serviços de atendimentos médicos, desenvolvendo um índice de Atraso Consciente da Tolerância (TAD) que incorporava explicitamente as informações de tolerância do usuário na avaliação de atrasos e comparava com a prática atual, por meio de dados sintéticos e dados hospitalares reais (Wang *et al.,* 2024). No mesmo sentido, um estudo desenvolveu dados sintéticos por meio da técnica de *Ordered Correlation Forest*, que lida com a não linearidade dos dados, para identificar autoavaliações de bem-estar subjetivo e autoavaliações em domínios da saúde de indivíduos, também contribuindo com a privacidade e alguns concernimentos dos pacientes (Zaman *et al.,* 2021).

Relacionado ao uso de técnicas de aprendizado de máquina para administração e simulação de bancos de dados, um estudo aplicou a técnica de *Machine Learning* chamada de *Autocodificadores Variacionais* (VAEs), ao qual por meio de amostras pequenas, grandes amostras são geradas e desenvolvem agentes não encontrados nos dados da amostra, mas que apresentam o risco de criar agentes inexistentes de uma população real utilizada como base (Sané *et al.,* 2025). Outro trabalho utilizou a técnica de *Microsimulação Espacial*, que cria populações sintéticas de indivíduos ou unidades em um nível micro (por exemplo, pessoas, famílias, empresas) e as localiza em um espaço geográfico, replicando as características estatísticas da população real e sua distribuição espacial, para subsidiar a tomada de decisões locais, referente ao comportamento alimentar e suas desigualdades associadas aos bairros carentes do Reino Unido, utilizado posteriormente para incentivar uma alimentação mais saudável de moradores (Schwaller *et al.,* 2021).

Um dos artigos indicados apresentou o uso da aplicação de uma abordagem de aprendizagem de máquina não-supervisionada para criar dados sintéticos e avaliar se questionários de aplicação destes dados, funcionavam na ausência de invariância de mensuração, condição necessária para analisar diferenças nos padrões de resposta em grupos diferentes de uma pesquisa, como gênero, idade e outros (Hahn-Klimroth *et al.,* 2024). Por fim, outra pesquisa utilizou da criação de dados sintéticos para testar dois algoritmos propostos para identificar a noção de causalidade, por meio de Modelos Vetoriais Autorregressivos (VAR), que são usados para modelar a evolução temporal de múltiplas séries temporais inter-relacionadas, assumindo que o valor atual de cada série temporal depende de seus próprios valores passados e dos valores passados de outras séries no sistema, aplicados em cenários estáticos e dinâmicos de teste aos algoritmos (Di Francesco, 2025). A Tabela 1 mostra os resultados desta busca.

Tabela 1 – Resultados nas Ciências Sociais Aplicadas

Autor/Ano	Título do Estudo	Área do Conhecimento	Objetivo do Estudo	Contexto da Utilização	Método de Geração de Dados Sintéticos
Eman et al. (2019)	Estudo sobre privacidade de dados na COVID-19	Ciências da Saúde	Preservar identidade em estudos de saúde	Pacientes em testes da COVID-19	Geração de dados sintéticos com LLM
Wang et al. (2024)	Preferências por vacinas COVID-19	Ciências Sociais	Simular respostas sobre preferências vacinais	Estudo empírico sobre vacinas	Modelos LLM com validação conjunta
Shortt et al. (2005)	Planejamento de atendimento regional	Ciências da Saúde	Distribuição de médicos generalistas	Atendimento em áreas remotas do Reino Unido	Matriz de Dados Sintéticos (MDS)
Wang et al. (2024)	Sequenciamento de atendimentos médicos	Saúde Pública	Avaliar tolerância de atraso em pacientes	Agendamento hospitalar	Dados sintéticos com modelo TAD
Zaman et al. (2021)	Bem-estar subjetivo	Psicologia Social	Modelar avaliações subjetivas com privacidade	Autoavaliação em domínios da saúde	Ordered Correlation Forest
Sané et al. (2025)	Geração de grandes amostras por VAEs	Ciência da Computação	Simular agentes populacionais	Bases de dados com amostras pequenas	Autocodificadores Variacionais
Schwaller et al. (2021)	Comportamento alimentar geolocalizado	Ciências Sociais Aplicadas	Estudar alimentação e desigualdade regional	Moradores de bairros carentes do Reino Unido	Microsimulação Espacial
Hahn-Klimroth et al. (2024)	Invariância de mensuração em questionários	Psicometria	Verificar eficácia de questionários sem invariância	Grupos distintos por idade/gênero	Aprendizado não supervisionado
Di Francesco (2025)	Causalidade com VAR e dados sintéticos	Estatística	Testar algoritmos de causalidade em VAR	Cenários dinâmicos de múltiplas séries	Modelos Vetoriais Autorregressivos (VAR)

Fonte: Elaborado pelos autores (2025)

Essa busca inicial permitiu uma visão mais detalhada do uso de dados sintéticos nos estudos de ciências sociais aplicada, que se mostrou mais relacionado aos processos de administração de dados e de criação de dados sintéticos, utilizando de formas diversas de aprendizado de máquina, mas aplicadas na grande maioria em áreas variadas, principalmente das ciências humanas. Se é possível considerar pacientes como consumidores, esses resultados são o mais próximos de se chegar a uma ideia de uso dos dados sintéticos no campo do consumo. Em uma outra visão mais pessimistas, estes resultados iniciais ligaram um alerta

sobre a possível escassez de artigos no campo do comportamento do consumidor e do marketing.

Na segunda etapa, utilizou-se as mesmas bases de dados, aplicando os termos "*synthetic data*" AND "*consumer behavior*" "OR" *marketing*" OR "*consumer decision-making*", também pesquisados no campo de "Tópicos" (Título, Resumo e Palavras-chaves) e considerando os mesmos filtros (Área de Ciências Sociais Aplicadas, Revisado por Pares, Tipo de Documento e Área de Conhecimento) da primeira etapa, para identificar de forma mais aprofundada o uso de dados sintéticos no campo do comportamento do consumidor e do marketing.

A pesquisa retornou apenas 2 resultados, que também foram acessados de forma integral. O primeiro resultado, uma pesquisa que simulava transações de consumidores por meio de dados sintéticos gerados por redes adversárias generativas (GANs), utilizadas para simular transações sob restrições de estoque, sendo pioneiro em um sistema experimental engenhoso com implicações práticas para a operação e estratégia de varejo do mundo real (Tkachuk *et al.,* 2024).

O segundo artigo se refere a uma revisão bibliográfica que buscava familiarizar pesquisadores e profissionais com novas fontes de dados e técnicas de análise para estudar o comportamento do consumidor em escala, incluindo, o uso de dados sintéticos para alavancagem de dados no contexto do *Big Data* para o consumo (Chang & Mukherjee, 2023). Novamente, nenhum dos artigos encontrados aplicavam o uso de dados para uso em pesquisas acadêmicas, sendo uma delas utilizada para simulações de mercado e outro uma revisão do uso de dados do consumo em escala, ambos feitos por periódicos da ciência da computação. A tabela 2 mostra os resultados desta busca.

Tabela 2 – Resultados nas áreas do Comportamento do Consumidor e do Marketing

Autor/Ano	Título do Estudo	Área do Conhecimento	Objetivo do Estudo	Contexto da Utilização	Método de Geração de Dados Sintéticos
Tkachuk et al. (2024)	Simulação de transações de consumo com GANs	Marketing	Simular comportamento de compra sob restrições de estoque	Varejo e estratégia operacional	Redes Adversariais Generativas (GANs)
Chang & Mukherjee (2023)	Uso de dados sintéticos no Big Data para consumo	Comportamento do Consumidor	Explorar novas fontes de dados para entender o consumo	Escalabilidade no estudo de consumo	Apresentação teórica sobre dados sintéticos

Fonte: Elaborado pelos autores (2025)

A análise qualitativa dos artigos também foi feita por meio da linguagem de programação Python e o uso do software Anaconda, utilizando a biblioteca '*nltk'* (Processamento de Linguagem Natural) para limpar, *tokenizar*, remover *stopwords* e lematizar, buscando normalizar o conteúdo textual. Após, analisou-se a frequência de palavras com a biblioteca *'collection'* e com módulo *'Counter*', buscando contar repetições das palavras, e uma Análise de Tópicos com LDA (*Latent Dirichlet Allocation*), por meio da biblioteca '*scikit-learn'* e do módulo '*LatentDirichletAllocation',* para classificar essas palavras em tópicos. Por fim, foi realizado um agrupamento por similaridade de conteúdo por meio de KMeans, utilizando o módulo *'KMeans'* da biblioteca '*scikit-learn'.*

Os textos completos dos artigos foram importados nas duas pesquisas ("Ciências Sociais Aplicada" e "Marketing OU Comportamento do Consumidor") e foram criadas categorias e subcategorias de análise com base nos objetivos, técnicas de geração de dados e tipos de aplicação nos campos, resultando em códigos temáticos como:

(1) Dados Sintéticos em Pesquisas;
(2) Técnicas de Simulação de Dados;
(3) Pesquisas Simuladas;
(4) Modelagem de Dados em Pesquisa Simuladas;
(5) Técnicas de Aprendizado de Máquina em Pesquisas

Após, foi utilizada uma análise quantitativa dos dados extraídos também por meio de Python, por meio da biblioteca '*panda's* para manipulação dos dados. Inicialmente, os artigos incluídos foram organizados em uma planilha estruturada contendo as variáveis autor/ano, título do estudo, área do conhecimento, objetivo do estudo, contexto da utilização e método de geração de dados sintéticos e acrescentado os códigos temáticos identificados na pesquisa qualitativa.

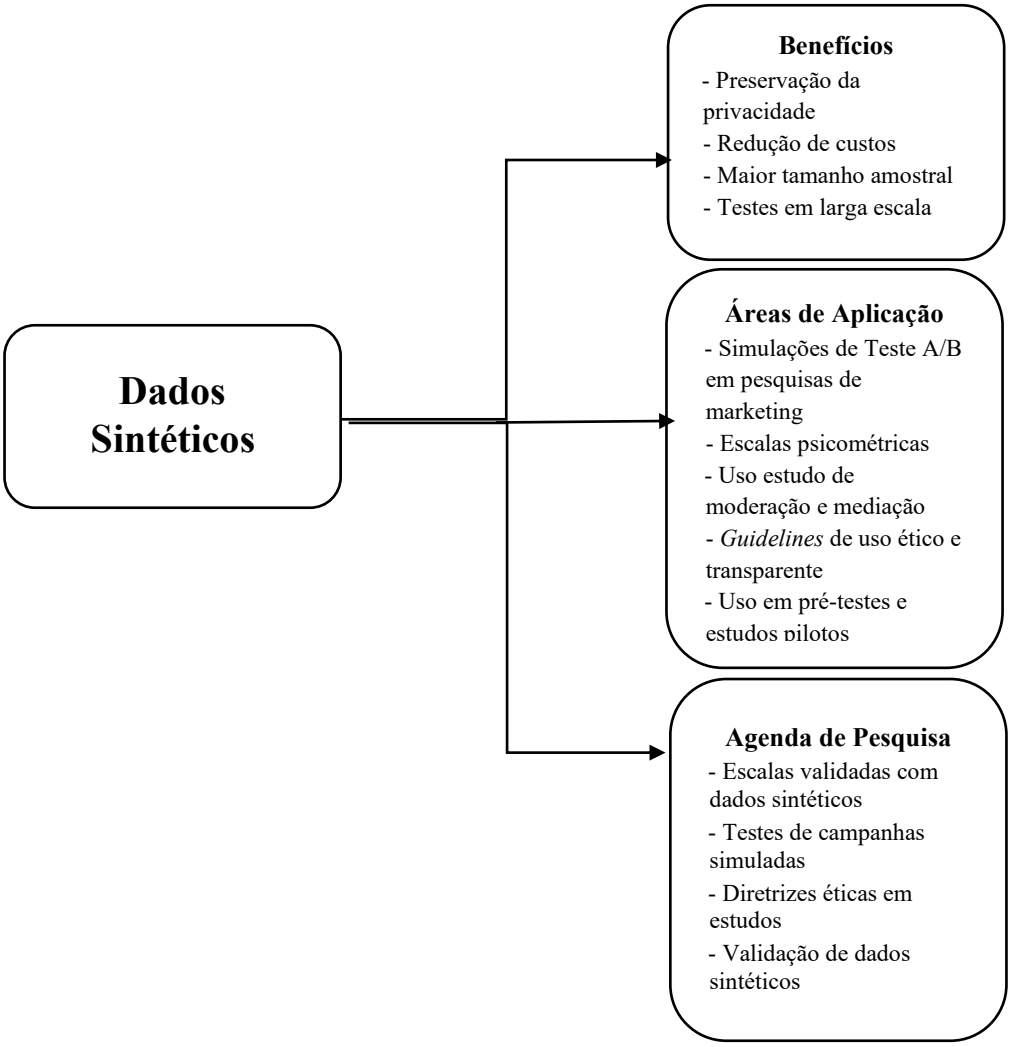
Dessa forma, realizaram-se estatísticas de tabulação cruzada entre os temas e gráficos exploratórios que permitiram identificar a frequência de uso das metodologias de geração de dados sintéticos, a conexão entre os campos pesquisados e os campos mais recorrentes de aplicação, o que contribuiu com a informações obre o estado atual da pesquisa, desafios e lacunas identificadas, direções para pesquisas futuras.

DESAFIOS, LACUNAS IDENTIFICADAS, DIREÇÕES PARA PESQUISAS FUTURAS E BENEFÍCIOS PREVISTOS DE USO DOS DADOS SINTÉTICOS

A literatura mostrou-se ainda incipiente e dispersa, com a maior parte dos estudos se concentrando fora do campo do marketing e do comportamento do consumidor, com predominância em áreas como saúde, atendimento, alimentação e questões sociais, além de abordagens técnicas e computacionais, com pouco aprofundamento em aspectos comportamentais.

Os resultados apontam que estudos futuros poderiam se aprofundar no uso de dados sintéticos em diversas áreas de aplicação, com benefícios significativos de uso e em campos diversos que garantem uma agenda de pesquisa condizente com a realidade nas áreas do comportamento do consumidor e do marketing. A seguir, um *framework* apresenta as principais perspectivas futuras do uso de dados sintéticos no campo do comportamento do consumidor e do marketing.

FRAMEWORK DE PERSPECTIVAS FUTURAS DO USO DE DADOS SINTÉTICOS NAS ÁREAS DO COMPORTAMENTO DO CONSUMIDOR E DO MARKETING



Fonte: Elaborado pelos autores (2025)

Os resultados mostram que no campo das ciências sociais aplicadas, a utilização de modelos generativos por meio de GANs (*Generative Adversarial Networks*) ainda é a mais utilizada, apesar de novas técnicas terem sido aprofundadas, como *Ordered Correlation Forest* (Zaman et al., 2021), Autocodificadores Variacionais (Sané *et al.,* 2025) e Microsimulação Espacial (Schwaller *et al.,* 2021), para simulação de perfis e aumento dos bancos de dados para análises mais profundas. De forma mais direta, os estudos no campo se concentram mais nos processos de administração de dados e de criação de dados sintéticos, e pouco em análises que simulem comportamentos, atitudes e opiniões para encontrar resultados mais expressivos.

No campo do marketing e do comportamento do consumidor, os resultados focados em simular transações sob restrições de estoque e da familiarização do uso de dados sintéticos para uso em pesquisas do comportamento do consumidor em escala, serviram para compreender que o temas ainda é novo e precisa de melhor aprofundamento e principalmente, iniciativa.

As contribuições emergentes identificadas mostram que o uso de dados sintéticos como substituto de dados sensíveis pode ser um caminho interessante para o campo do comportamento do consumidor e do marketing, principalmente em pesquisas que requerem dados de comportamento real, mas que possuem forte restrição ética ou legal, como a LGPD. Outrossim, o uso dos dados sintéticos traria um aumento significativo na base de dados das pesquisas, constantemente limitada em estudos experimentais, o que viabilizaria análises estatísticas robustas, com mais variáveis e cenários. Por fim, a possível redução de custos em pré-testes e estudos pilotos de pesquisa, permitiria testes em larga escala sem a necessidade de coleta primária real, o que seria de muito valia pela atual conjuntura do pouco ou quase nulo apoio financeiro para pesquisas acadêmicas.

As lacunas identificadas poderiam compreender uma relação considerável de condições a se relatar, porém, este estudo decide foca em apresentar algumas lacunas iniciais e que serviriam de base para possíveis interessados no desenvolvimento de pesquisas com dados sintéticos no campo do comportamento do consumidor e do marketing. Por exemplo, a baixa aplicação em pesquisas comportamentais mostra o quanto este campo pode ser ainda explorado, pois, a maioria das aplicações atuais está em áreas técnicas, em estudos que apresentam algumas situações superficiais do campo das ciências sociais aplicadas, sendo ainda mais escassas em relação às pesquisas acadêmicas do comportamento do consumidor e do marketing.

Percebe-se ainda a falta de experimentos de campo validados por dados reais, pois, a maioria dos estudos possui validações que ocorrem com dados simulados que são comparados com o próprio modelo. Além disso, os experimentos poderiam dar maior atenção no uso de dados sintéticos em estudos pilotos e em pré-testes, ao qual uma base anterior de outros estudos, poderia ser utilizada na geração de dados sintéticos por meio da técnica de GANs (Generative Adversarial Networks). Um pouco mais adiante e de forma complementar as lacunas identificadas, os estudos poderiam se adiantar e se atentar ao desenvolvimento de diretrizes éticas e de validação dos dados sintéticos, promovendo *frameworks* padronizados para avaliar a fidedignidade dos dados sintéticos usados. Essa ação poderia contribuir com a quebra de barreiras dos pesquisadores que não consideram ainda o uso dos dados sintéticos, principalmente, por não perceber um arcabouço técnico e literário que sustente e incentive o uso em seus estudos.

Estudos futuros poderiam utilizar dados sintéticos para estudos de caso em pesquisas acadêmicas de marketing, permitindo comparação entre campanhas reais e campanhas simuladas e decisões comportamentais do consumidor. Por exemplo, identificar se campanhas reais de marca e modelos alternativos com variações, no estilo Teste A/B, poderiam apresentar resultados diferentes em relação as variáveis testadas, como satisfação, engajamento e outras. Estudos também poderiam explorar o uso de dados sintéticos para simular respostas no desenvolvimento de escalas comportamentais, criando bancos de dados sintéticos a partir de descrições de consumidores reais ou dados de painéis. Por meio de informações criadas utilizando modelos generativos, pesquisadores poderiam testar ajustes de escala de forma mais intensa e mais rápida, antes de levar uma versão definitiva para análises exploratórias e confirmatórias.

Estudos focados em mediação e moderação também poderiam absorver as vantagens dos dados sintéticos. Por meio de uma base inicial de análise de causalidade, pesquisadores podem explorar estes resultados para desenvolver bases consequentes, no ensejo de testar variáveis mediadoras e moderadoras antes de experimentos reais, o que não fugiria da proposta do

formato *Bootstrapping* aplicado atualmente. Com isso, muitos modelos de interação poderiam ser testados previamente, evitando o desperdício de tempo e dinheiro em aplicações de experimentos reais que perdem resultados palpáveis, devido ao uso inadequado de variáveis ou designs experimentais mal elaborados.

Pesquisas poderiam desenvolver *guidelines* para o uso ético e transparente em pesquisas com humanos simulados, buscando evitar vieses e assegurar validade externa das simulações. Como o tema ainda é recente e pouco explorado, essas orientações serviriam para não acontecer abusos no uso de dados sintéticos, por exemplo, criações com base em informações irreais e/ou imprecisas dos indivíduos, validação insuficiente para prosseguir com uma versão de teste real e principalmente, o uso de dados sintéticos como se fossem dados reais. Esse último, liga até mesmo um grande alerta, de que os dados sintéticos devem servir de apoio e simulação, e não como versão definitiva e aplicada como real em estudos do campo. Essas *guidelines* dariam a orientação necessária, nos mesmos moldes que já acontecem em pesquisas no campo das ciências humanas.

Analisando de forma prática, um dos principais benefícios do uso de dados sintéticos seriam na contribuição do cruzamento de informações acadêmicas com informações reais do mercado. Por exemplo, estudos que buscam entender intenção de compra, satisfação com a marca e outros, poderiam utilizar dados reais de mercado, em pesquisas aplicadas, para identificar lacunas nestas teorias, o que promoveria uma condição favorável de aplicação em pesquisas acadêmicas reais, pois já estariam sustentadas pelos dados sintéticos criados por meio de informações que são advindas do mercado real. Adiante, estudos também poderiam comparar os resultados de dados sintéticos com dados reais em escalas tradicionais de marketing (por exemplo, *Purchase Intention, Net Promoter Score*).

O uso dos modelos generativos de dados sintéticos também pode contribuir com a criação de bases sintéticas de diferentes personas com perfis de consumo, permitindo testes e refinamento de campanhas reais de marketing. Por exemplo, os dados de resultado de estudos acadêmicos poderiam ser replicados por meio de dados sintéticos e aplicados em simulações de testes de campanhas reais de marcas, incluindo ações experimentais nos cenários digitais, para testar reações a estímulos de marca, anúncios, preços ou experiências de compra.

Um outro ponto que pode ser aplicado aos cenários reais seria o da criação de protocolos e diretrizes sobre o uso responsável de dados sintéticos em pesquisas com humanos simulados, incluindo o desenvolvimento de materiais didáticos e cursos para pesquisadores do comportamento do consumidor e do marketing utilizarem, promovendo a segurança e a eficácia. Este protocolo poderia ser útil para pesquisas acadêmicas e pesquisadores de mercado que encontram as mesmas condições limítrofes de participantes reais.

LIMITAÇÕES DO ESTUDO

A principal limitação do estudo é o próprio motivo de seu desenvolvimento. O baixo número de artigos nas pesquisas acadêmicas no campo do comportamento do consumidor e do marketing, mostrou que o tema ainda é um tabu em relação ao uso das técnicas de modelos generativos. Dessa forma, para aumentar o arcabouço teórico, optou-se por também incluir pesquisas no campo das ciências sociais aplicadas, o que contribuiu de forma limitada aos estudos, principalmente, por se tratar mais da área da saúde, alimentação e atendimento hospitalar, mesmo que a contribuição tenha apresentado novas técnicas de geração de dados sintéticos.

Outra limitação se refere a falta de validação realística, com pouca verificação da fidelidade comportamental dos dados sintéticos em relação aos dados reais. Os estudos

apresentam algumas situações de replicação de dados para compreender algumas condições na saúde e alimentação, mas não avançam para pesquisas reais e comparações em relação a essa fidelidade dos dados sintéticos.

Um outro ponto percebido se refere a pouca integração com escalas psicométricas, comumente utilizadas em estudos do comportamento do consumidor, apresentando uma falta de testes dos dados sintéticos que pudessem simular respostas em escalas como intenção de compra, engajamento, satisfação e outros, fatores que podem causar insegurança quanto à reprodutibilidade, validade externa e integridade científica.

CONCLUSÕES

A literatura sobre dados sintéticos em comportamento do consumidor e marketing mostrou-se limitada e ainda não emergente, apesar das vastas oportunidades percebida, principalmente, em relação a simulações que precedem pesquisas reais e consideram as questões de privacidade dos participantes. Poucos artigos exploram diretamente o tema, mas percebe-se um crescente interesse em métodos de inteligência artificial aplicados ao desenvolvimento de dados sintéticos nas situações que se aproximam do consumo, sejam relacionados a formas de atendimento, como no caso de pacientes da saúde, no interesse da alimentação de pessoas, na gestão de bancos de dados relacionados ao marketing, na oportunidades de identificar tendências comportamentais, no cuidado com a privacidade de indivíduos ou simplesmente, em técnicas que contribuem com a escassez de dados.

As principais lacunas envolvem baixa aplicação em pesquisas comportamentais, ausência de experimentos de campo validados por dados reais, exploração dos dados sintéticos em estudos pilotos e em pré-testes e promoção de frameworks padronizados que contribuam com pesquisadores interessados no uso dos modelos generativos de dados sintéticos. Sugere-se uma agenda futura que inclua *guidelines* éticos, simulações de comportamento de consumo, uso em testes A/B e pesquisas experimentais com mediação e moderação. O ponto mais importante é compreender que áreas como saúde e finanças, já se encontram mais avançadas em relação ao uso de dados sintéticos, o que poderia fomentar o interesse nas pesquisas no campo das ciências sociais aplicadas, incluindo em pesquisas acadêmicas do comportamento do consumidor e do marketing.

REFERÊNCIAS

Ceipek, R. (2019). Technological diversification: a systematic review of antecedents, outcomes and moderators. *International Journal of Management Reviews*, pp. 1-32.

Chang, H. & Mukherjee, A. (2023). Machine Learning and Consumer Data. *Cornell University*.

D'Amico *et al.* (2023). Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology. *Clinical Cancer Informatics*.

Di Francesco, R. (2025). Ordered correlation forest. *Econometric Reviews.*

Dwivedi, S. (2024). Synthetic data and European General Data Protection Regulation: *Ethics, quality and legality of data sharing*. 6(4), 332.

El Emam, K., Mosquera, L., & Bass, J. (2020). Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. *J Med Internet Res*, 22(11).

Eno, J., & Thompson, C. (2008). Generating Synthetic Data to Match Data Mining Patterns. *IEEE Internet Computing*, 12(3), 78-82.

Galvagno, M., & Dalli, D. (2014). Theory of value co-creation: a systematic literature review. *Managing Service Quality: An International Journal*, 24(6), 643-683.

Goodfellow, I.J.; Pouget-Abadie, J., Mirza, M. *et al.* (2015). Generative adversarial nets, *Cornell University.*

Grewal, D.; Satornino, C.; Davenport, T. & Guha, A. (2024). How generative AI Is shaping the future of marketing, *Cornell University.*

Hahn-Klimroth, M., Dierkes, P. W., & Kleespies, M. W. (2024). An Unsupervised Learning Approach to Evaluate Questionnaire Data—What One Can Learn from Violations of Measurement Invariance. *Data Science Journal*.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical Mediation Analysis in the New Millennium. *Communication Monographs*, 76(4), 408-420.

Hermann, E. & Puntoni, S. (2024). Artificial intelligence and consumer behavior: From predictive to generative AI. *Journal of Business Research*.

Jennifer Taub, Mark Elliot, Joseph W. Sakshaug (2026). A Study of the Impact of Synthetic Data Generation Techniques on Data Utility using the 1991 UK Samples of Anonymised Records, *The University of Manchester.*

Kokosi, T. & Harron, K (2022). Synthetic data in medical research. *BMJ Medicine*.

Le Cun, Y.; Bengio, Y.; Hinton, G. (2015). Deep learning. Nature, 521 (436-444).

Niamh, K. S., Moore, A., Coombes, M., & Wymer, C. (2005). Defining regions for locality health care planning: a multidimensional approach. *Social Science & Medicine*, 60(12), 2715-2727.

Potluru, V. K. et al. (2023). Synthetic Data Applications in Finance. *Cornell University*.

Rankin, D., Black, M., Bond, R., Wallace, J., Mulvenna, M., & Epelde, G. (2020). Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. *JMIR Med Inform*, 8(7).

Richard Timpone, Yongwei Yang. (2024) Artificial Data, Real Insights: Evaluating Opportunities and Risks of Expanding the Data Ecosystem with Synthetic Data. *International Conference for Computational Social Science.*

Sané, A. R., Vandanjon, P. O., Belaroussi, R., et al. (2025). A comprehensive investigation of variational auto-encoders for population synthesis. *J Comput Soc Sc*, 8 (13).

Schwaller, E., et al. (2021). Area inequalities in fruit and vegetable intake in England: a spatial microsimulation, cross-sectional study. *The Lancet*, 398, S78.

Tkachuk, S., Łukasik, S. & Wróblewska, A. (2024). Consumer Transactions Simulation through Generative Adversarial Networks. *Cornell University*.

Wang, M., Zhang, D. J., & Zhang, H. (2024). Large Language Models for Market Research: A Data-augmentation Approach. *Cornell University.*

Wang, S., Li, J., Ang, M., & Ng, T. S. (2024). Appointment Scheduling with Delay Tolerance Heterogeneity. *INFORMS Journal on Computing*, 36(5), 1201-1224.

Zaman, B., Ramos, L. M. L., Romero, D., & Beferull-Lozano, B. (2021). Online Topology Identification From Vector Autoregressive Time Series. *IEEE Transactions on Signal Processing*, 69, 210-225.