ORIGINAL ARTICLE

Optimized Sequential Synthesis of Tabular Microdata

Cameron D. Bale[1] | **Harrison Quick**[2]

[1]Decision Sciences, Drexel University, Philadelphia, PA 19104, USA

[2]Department of Epidemiology and Biostatistics, Drexel University, Philadelphia, PA 19104, USA

Correspondence
Cameron D. Bale, 3220 Market Street, Office 734, Drexel University, Philadelphia, PA 19104, USA
Email: cdb327@drexel.edu

Funding information
None

Data synthesis is a promising approach for generating protected data sets, especially in the face of increasing privacy concerns and legislation. However, the costs and expert knowledge required to implement and tune data synthesis models may prevent data owners from using synthesis as a disclosure limitation solution. In this paper, we study legal privacy criteria used to evaluate synthetic data and show that these criteria are met in expectation when the synthetic data is sampled from the same distribution as the confidential data. We propose a new sequential synthesis method that utilizes Bayesian optimization to tune synthesis models to approximate the data generating distribution. We also implement an algorithm that pairs with the proposed synthesis approach to reduce the probability of attribute disclosure for a sensitive binary variable. In an empirical application to South Korean COVID-19 Location data, we find that the proposed synthesis method produces synthetic data with comparable privacy as measured by the legal criteria, reduced attribute disclosure risk, and better utility than a baseline method known to produce legally anonymous synthetic data.

KEYWORDS
Working keywords.

1 | INTRODUCTION

Data owners that collect and disseminate confidential data must maintain the utility of their data for analyses while protecting the identities and sensitive information of data subjects. Furthermore, to comply with privacy law, data owners can either implement privacy and security measures for confidential data (failing to do so can result in significant fines - for example, see Koch (2018)), or anonymize personal data (European Data Protection Board, 2020). Anonymization methods produce protected data that falls outside of the scope of privacy legislation. Recent privacy legislation (European Data Protection Board, 2020; Working Party, 2014), and related research (Bale et al., 2022) highlight the need for robust anonymization methods that produce legally anonymous data.

Both academic researchers (e.g., (Quick, 2021)) and practitioners have shown the potential for synthetic data to provide a desirable balance between data utility and privacy. A recent report published by the Joint Research Centre (JRC) of the European Commission studied the use of synthetic data for informing policy in a privacy-preserving manner (Hradec et al., 2022), claiming that, "This artificially generated data is highly representative, yet completely anonymous. As it does not contain any one-to-one relationships to actual data subjects, the risk of re-identification is effectively eliminated," (as long as the synthetic data meets certain privacy criteria) (pg. 44). The company that performed the data synthesis, MOSTLY.AI, states that their synthetic data is exempt from data privacy legislation[1]. Others make similar claims, such as Hintze and El Emam (2020) who discuss how synthetic data can be used as a solution to transferring sensitive data from the EU to the US after the invalidation of the EU-U.S. Privacy Shield agreement. However, choosing a data synthesis model to balance the utility and privacy for a given data set is not a trivial task.

To produce synthetic data, a data owner uses a generative model to capture the data generating process of the confidential data $\mathbf{Y}$. This generative model can take many forms, ranging from deep learning models such as generative adversarial networks (Lee and Anand, 2020), to Bayesian models from which the synthetic data is sampled from the posterior predictive distribution (Quick, 2021). Regardless of the synthesis model, synthetic data must meet privacy requirements and maintain the utility of the original data to be considered useful.

Some synthesis methods produce differentially private data which is suited for specific use cases but can reduce utility significantly in others (Bowen and Snoke, 2021). We focus on non-differentially private synthesis methods (hereafter referred to as *synthesis methods*) which can produce synthetic data with greater utility but require manual tuning and privacy risk assessment to ensure that the synthetic data is safe to release. Open source data synthesis solutions such as `synthpop` (Nowok et al., 2016) enable data owners to train synthesis models and evaluate the utility of synthetic data, but data owners must still manually adjust model parameters to balance utility and privacy. Even in cases where synthesis models are expertly tuned to balance utility and privacy, the academic literature has not examined whether the data sets that are produced can be considered legally anonymous and therefore fall outside the scope of privacy law. Commercial solutions which *do* claim legal anonymity either cost money or place limits on the amount of data which can synthesized at any given time. These commercial solutions are often based on deep learning models which can require large amounts of sensitive training data to produce synthetic data with good utility (Hradec et al., 2022). Overall, the complexity of tuning synthesis models and the technological and cost barriers to legal data synthesis can prevent data owners from fully reaping the benefits of using data synthesis as a privacy solution.

We reduce these barriers by proposing a general sequential synthesis method which produces legally anonymous synthetic data based on the privacy criteria used by MOSTLY.AI in the JRC report (Hradec et al., 2022). We augment our proposed synthesis method with an algorithm that samples synthetic records specifically to reduce the probability

[1] https://mostly.ai/privacy-and-security/

of attribute disclosure for a sensitive binary variable. First, we outline the legal privacy criteria mathematically and prove that the criteria are guaranteed to be met in expectation when synthetic data is drawn from the same data generating distribution as the sensitive data. Our proposed synthesis method incorporates commonly used synthesis models, such as multinomial logit models and classification and regression trees, and a Bayesian optimization process which automatically tunes the synthesis models to produce legally anonymous synthetic data. We utilize entirely open-source programming tools and all code can be found in our online repository. We verify the performance of our proposed method by comparing it to the data synthesis method from MOSTLY.AI (which is known to produce legally anonymous synthetic data) in an application to a data set containing marked locations of COVID-19 patients in South Korea. Our empirical results show that the proposed synthesis method produces synthetic data with comparable privacy to the MOSTLY.AI method based on the legal privacy criteria, increased protection against attribute disclosure, and better data utility.

The paper proceeds as follows. In Section 2, we provide a review of the relevant literature. In Section 3, we study the privacy measures used by MOSTLY.AI to determine whether synthetic data is legally anonymous, and show that synthetic data that is drawn from the same data generating distribution as the sensitive data is guaranteed to meet the privacy metrics in expectation. Section 4 describes a simulation study which we use to verify the privacy criteria results from Section 3. Section 5 describes our proposed synthesis method which is optimized to meet the criteria from Section 3. In Section 6 we apply our proposed synthesis method and the synthesis method from MOSTLY.AI to the location data of South Korean COVID-19 patients and compare the privacy and utility results of both synthesis methods. We conclude and discuss opportunities for further research in Section 8.

2 | LITERATURE REVIEW

The key idea behind our proposed methodology is that synthetic data meets the legal privacy criteria in expectation when it is sampled from a synthesis model that approximates the true data generating distribution, allowing us to view the confidential data and the synthetic data as two random samples from the same data generating distribution. The literature often assesses the utility of synthetic data using measures of distributional similarity between the synthetic and original data. Examples include the propensity score mean-squared error (pMSE) and pMSE ratio (Snoke et al., 2018), the Kolmogorov-Smirnov distance between propensity score distributions (Bowen et al., 2021), and differences between cross-tabulations of the synthetic and original variables (Schneider et al., 2023; Drechsler and Hu, 2021).

We focus on the pMSE ratio because it is widely used and gives a single valued measure of distributional similarity between two data sets. To compute the pMSE ratio we first train a discriminative model (common choices are a logistic regression or a classification and regression tree (CART)) to discriminate between synthetic and original records in a data set containing the combined synthetic and original data. The propensity score $\hat{p}_i$ of each record in the combined data is the in-sample estimate of the probability that a given record is synthetic. The propensity score mean-squared error (pMSE) is computed as

$$pMSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - c)^2, \tag{1}$$

where N is the number of combined synthetic and original records, and c is the fraction of the N records that are synthetic (typically $c = 0.5$). Snoke et al. (2018) derived the expected value of the pMSE when logistic regression is used as the propensity score model for the following two cases, (1) when the confidential data $\mathbf{Y}$ is treated as fixed

and the synthetic data **Z** is a matrix of random variables drawn from a synthesis model that correctly approximates the data generating distribution $\mathcal{D}$, and (2) when two synthetic data sets, both treated as matrices of random variables, are both drawn from the same data generating distribution or synthesis model. For the first case, the expected pMSE between **Y** and **Z** is

$$\mathbb{E}(pMSE) = (k-1)(1-c)^2 c/N, \tag{2}$$

where k is the number of variables that consist of synthesized values. We can compare the expected pMSE to the calculated pMSE using the pMSE ratio,

$$\mathcal{R} = pMSE/\mathbb{E}(pMSE), \tag{3}$$

where a value of $\mathcal{R} \approx 1$ indicates that the synthetic data set is properly synthesized. This approach to calculating the pMSE and pMSE ratio is commonly used in the literature (*e.g.*, (Guo and Hu, 2022; Schneider et al., 2023)).

2.1 | Differentially Private Data Synthesis

Traditional differential privacy is known to severely reduce data utility, especially for detailed microdata (Ruggles et al., 2019). Some work has combined differential privacy with data synthesis in hopes of producing synthetic data that meets a formal privacy definition and provides better utility than the traditional differential privacy approach. Bowen and Snoke (2021) performed a detailed analysis of the methods and data sets produced in the 2018-2019 National Institute of Standards and Technology Public Safety Communications Research (NIST PSCR) Division's "Differential Privacy Synthetic Data Challenge". These authors found that the synthetic data sets had relatively low utility as reported by measures of distributional similarity such as the pMSE ratio (Snoke et al., 2018). For example, in Match #2 for privacy parameter $\epsilon = 1$, the best average log pMSE ratio across synthetic data sets was 7.56 when using logistic regression as the discriminatory model. This means the average deviation in propensity scores was over 1900 times larger between the differentially private data and the original data than what would be expected if the data were correctly synthesized.

While the distributional similarity of differentially private synthetic data to the original data may be poor, differentially private synthetic data can preserve utility for specific use cases. Asghar et al. (2020) showed that a differentially private Gaussian copula model can generate synthetic mixed variable data sets that provide highly accurate answers to one-way and two-way positive conjunction queries. However, the utility of this data is limited to the queries it is designed to answer. In addition, a major assumption underlying a copula model is that the joint density of data samples is unimodal, which is often not the case.

Recent work combined differential privacy with data synthesis for Poisson-distributed count data (Quick, 2021). The authors used a Bayesian synthesis model with priors chosen to ensure the sampled data was differentially private. Quick (2022) adjusted this approach by truncating the prior distributions based on publicly available data. This reduced the amount of smoothing toward the prior distributions, thereby improving the utility of the synthetic data. In an application data on Pennsylvania cancer death counts, this method produced smaller variations between the real and synthetic death counts than comparable methods. For a relatively low privacy budget of $\epsilon = 4$, the largest difference was about thirty deaths, with most differences being less than ten deaths, and the synthetic data preserved the count

disparities between urban and rural areas and race/sex combinations. While this method offers a favorable balance between privacy and utility for synthetic count data, it cannot be applied to tabular microdata, which is the focus of this paper.

2.2 | Non-Differentially Private Data Synthesis

Bayesian models have been used to synthesize point-of-sale data Schneider et al. (2018), time series (Schneider and Abowd, 2015), location data with corresponding individual attributes Quick et al. (2015, 2018), and structured covariates that accompany online reviews (Schneider et al., 2023). These models often use a single parameter to control the privacy-utility trade-off via the prior distributions. For example, Schneider et al. (2018) used a privacy parameter κ as the inverse of the prior variance of the synthesis model fixed effects where larger (smaller) values of κ corresponded to weaker (stronger) privacy protection, and Schneider et al. (2023) used a privacy parameter λ as the inverse of the scale parameter of the laplace prior; larger values of λ smoothed the model coefficients towards zero and resulted in better privacy. In both cases, the final parameters had to be chosen by the data controller: Schneider et al. (2018) manually selected values for κ from 0.1 to 15 and (Schneider et al., 2023) performed a gridsearch for the value of λ which produced an average pMSE ratio closest to one.

CART models are also widely used for data synthesis, and have been shown to produce synthetic data with a desirable tradeoff between privacy and utility (Drechsler and Reiter, 2011; Reiter, 2005). CART models have been used to generate synthetic location data (Drechsler and Hu, 2021; Wang and Reiter, 2012) and have generated synthetic medical data for secondary analysis of clinical trial data (Azizi et al., 2021) and research on sensitive patient data sets (El Emam et al., 2021). The parameters of the CART models are either manually chosen or are left to default values that are known to perform reasonably well.

The approach proposed in this paper is very similar in principle to that of Jackson et al. (2022) who propose a synthesis method for large data sets based on saturated count models. These authors derive the expected values of risk and utility metrics for the synthetic data which could be set prior to data synthesis by the data controller using two parameters. This makes it easy for data controllers to define a desired level of risk and/or utility, and use the synthesis method to achieve that level. Similarly, we propose a method that optimizes the synthesis process such that risk metrics are met in expectation. This alleviates the need for data controllers to choose parameters for the synthesis model; instead, data controllers can choose the desired objective function with which to optimize the synthesis process.

3 | PRIVACY CRITERIA

Consider two confidential data sets, $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{N_1})^\mathsf{T}$ and $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_{N_2})^\mathsf{T}$, both of which contain N_1 records drawn independently from the same data generating distribution $\mathcal{D}$. A data owner is interested in using a data synthesis model to release a synthetic version of the data set $\mathbf{T}$, denoted $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{N_2})^\mathsf{T}$. Typically, $N_1 = N_2$, although this is not required. We refer to the data sets $\mathbf{Y}$, $\mathbf{H}$, and $\mathbf{Z}$ as the confidential, holdout, and synthetic data sets, respectively. For the synthetic data set to considered legally anonymous, it must meet the three privacy criteria used by MOSTLY.AI to evaluate the anonymity of their synthetic data, which we now define[2] (Hradec et al., 2022). The intuition behind the privacy criteria is that synthetic data provides adequate privacy if, on average, it is not more similar to the confidential data than a second confidential data set from the same data generating distribution, *i.e.*, the holdout

[2] We verified the written definitions of the criteria through email conversations with the company.

data set. We argue that the ideal data synthesis model is one which is identical to the data generating distribution $\mathcal{D}$, such that **Y**, **Z**, and **H** can be viewed as independent samples of records drawn from the same data generating distribution.

3.1 | Criterion One: The Identical Match Share

Let $d^z_{i,j} = \text{dist}(\mathbf{z}_i, \mathbf{y}_j)$ denote a distance measure (*e.g.*, Euclidean distance) between the ith and jth synthetic and confidential records, respectively. Further, let $d_i^{z(k)}$ denote the distance between the ith synthetic record and its kth nearest neighbor confidential record. The *distance to closest record* (DCR) is the distance between the ith synthetic record and its nearest neighbor confidential record, denoted $d_i^{z(1)}$. The *identical match share* (IMS) for the synthetic data is defined as

$$IMS_{\mathbf{Z}} = \frac{1}{N_2} \sum_{i=1}^{N_2} I(d_i^{z(1)} \leq \delta_{IMS}), \tag{4}$$

where $I(.)$ is the indicator function. The IMS is the proportion of synthetic records which are identical to their nearest neighbor training record. For records with continuous attributes, there is zero probability of having an exactly identical record in a finite population. In this case a small, non-zero δ_{IMS} may be specified as the threshold for records to be considered identical. Alternatively, continuous attributes may be discretized into bins, and a measure such as the Hamming distance may be used, in which case δ_{IMS} may be set to zero. Note that $IMS_{\mathbf{H}}$ can be defined similarly for the holdout data set. The first privacy criterion requires that the IMS for the synthetic data set is no more than the IMS for the holdout data set:

$$IMS_{\mathbf{Z}} \leq IMS_{\mathbf{H}}. \tag{5}$$

The ideal data synthesis model would be identical to $\mathcal{D}$ such that sampling synthetic records from the model is no more likely to produce a copy of a confidential record than sampling additional confidential records from $\mathcal{D}$. Privacy issues occur if a synthesis model generates a higher proportion of identical synthetic records than what would be produced from randomly sampling additional confidential records. This can occur if a synthesis model is overfit to the confidential data set and can reveal the presence of specific records. Synthetic data from such a model does not provide the plausible deniability that any similarity between a synthetic record and a confidential record is just as likely to occur between a holdout record and a confidential record (Platzer and Reutterer, 2021).

3.2 | Criterion Two: The Distance to Closest Record

We let $\mathbf{F}^{DCR(z)}$ denote the empirical cumulative distribution function of the DCRs from the synthetic data set. We denote the fifth percentile of $\mathbf{F}^{DCR(z)}$ as $p_{0.05}^{DCR(z)}$. Note that $\mathbf{F}^{DCR(h)}$ and $p_{0.05}^{DCR(h)}$ can be defined similarly for the holdout data set. The second privacy criterion requires that the fifth percentile of $\mathbf{F}^{DCR(z)}$ be no less than the fifth percentile of $\mathbf{F}^{DCR(h)}$:

$$p_{0.05}^{DCR(z)} \geq p_{0.05}^{DCR(h)}. \tag{6}$$

The ideal synthesis model would be identical to $\mathcal{D}$ to ensure that there is not a systematic tendency for the synthesis model to produce synthetic records which are more similar to confidential records than the holdout records to the confidential records.

3.3 | Criterion Three: The Nearest Neighbor Distance Ratio

Let the *nearest neighbor distance ratio* (NNDR) for the ith synthetic record be denoted as

$$NNDR_i^z = \frac{d_i^{z(1)}}{d_i^{z(5)}}, \tag{7}$$

which is bounded between $[0, 1]$ and denotes the ratio of the distances between the ith synthetic record and its nearest and fifth nearest neighbor confidential records. A ratio close to one indicates that the synthetic record is roughly equally similar to its nearest five confidential records. A ratio close to zero indicates that a synthetic record has disproportionately high similarity to its nearest neighbor confidential record than its fifth nearest neighbor confidential record. This could indicate that a synthetic record is revealing the presence of an outlier in the confidential data. However, this is acceptable if the holdout data provides similar information on outliers, *i.e.*, if the distributions of the nearest neighbor distance ratios are similar. Let $\mathbf{F}^{NNDR(z)}$ denote the empirical cumulative distribution function of the NNDRs of the synthetic data set and let $p_{0.05}^{NNDR(z)}$ denote the fifth percentile of this distribution. Note that $\mathbf{F}^{NNDR(h)}$ and $p_{0.05}^{NNDR(h)}$ can be defined similarly for the holdout data set. The third privacy criterion requires that the fifth percentile of the nearest neighbor distance ratios for synthetic records cannot be smaller than the fifth percentile for holdout records:

$$p_{NNDR}^z \geq p_{NNDR}^h. \tag{8}$$

When compared to the holdout data, the ideal data synthesis model will produce synthetic records that are not overly similar to outliers compared to the other confidential records.

3.4 | Satisfying the Privacy Criteria in Expectation

Now we assume that the synthetic data set $\mathbf{Z}$ is sampled from a data synthesis model that captures the data generating distribution $\mathcal{D}$. For a given synthetic record $\mathbf{z}_i$, let $C_{i,j}$ denote a Bernoulli random variable where $C_{i,j} = 1$ when the jth confidential record is identical to $\mathbf{z}_i$. Then the probability that $C_{i,j} = 1$ is given by $p(C_{i,j} = 1) = p(d_{i,j}^z \leq \delta) = p_C$. There are N_1 confidential records, which we assume are drawn independently and with replacement from $\mathcal{D}$. The random variable denoting the number of confidential records which are identical to $\mathbf{z}_i$ is given by $M_i = \sum_{j=1}^{N} C_{i,j}$, such that M_i is binomially distributed with parameters N_1 and p_C,

$$
\begin{aligned}
M_i &\sim Bin(N_1, p_C), \\
p(M_i = m_i) &= \binom{N_1}{m_i} p_C^{m_i} (1 - p_C)^{N_1 - m_i}.
\end{aligned}
\tag{9}
$$

The expected number of confidential records that are identical to a given synthetic record is $N_1 \times p_C$, which scales with the number of confidential records. Intuitively, the more records are in the confidential data set, the more likely there is to be an identical match between a confidential record and a given synthetic record.

To calculate the expected value of the identical match share when $\mathbf{Z}$ and $\mathbf{Y}$ are samples from $\mathcal{D}$, we first calculate the probability that there is at least one confidential record that is identical to a synthetic record, *i.e.*, for a given synthetic record, we are interested in

$$
\begin{aligned}
P(M_i \geq 1) &= 1 - P(M_i = 0) \\
&= 1 - \prod_{j=1}^{N} P(C_{i,j} = 0) \\
&= 1 - (1 - P_C)^N \\
&= p_{IMS}.
\end{aligned}
\tag{10}
$$

We define another Bernoulli random variable $\mathcal{M}_i = 1$ with probability p_{IMS} if the ith synthetic record has at least one match in the confidential data, and $\mathcal{M}_i = 0$ with probability $1 - p_{IMS}$. Then the random variable $\mathcal{M} = \mathcal{M}_1 + \mathcal{M}_2 + \cdots + \mathcal{M}_{N_2}$ is the total number of synthetic records which are identical to any given confidential record and is Binomially distributed with expected value $N_2 \times p_{IMS}$. The identical match share from (4) is calculated as

$$
IMS_{\mathbf{Z}} = \frac{1}{N_2} \mathcal{M}. \tag{11}
$$

The expected value of $IMS_{\mathbf{Z}}$ is given by

$$
\begin{aligned}
\mathbb{E}[IMS_{\mathbf{Z}}] &= \frac{1}{N_2} \mathbb{E}[\mathcal{M}] \\
&= \frac{1}{N_2} N_2 \times p_{IMS} \\
&= p_{IMS}.
\end{aligned}
\tag{12}
$$

Since the holdout and synthetic data sets are both identically drawn from $\mathcal{D}$, the above result holds for both data sets. The expected values of the identical match shares for the synthetic and holdout data sets are equal to p_{IMS}, therefore the identical match share criterion in (5) is met in expectation for synthetic data sets drawn from the same data generating distribution as the confidential and holdout data sets. Similar reasoning can conclude that the criteria for distance to closest record and nearest neighbor distance ratio will also be met in expectation for a synthetic data set drawn from the same data generating distribution $\mathcal{D}$.

3.5 | Attribute Disclosure

Record linkage, or re-identification disclosure attacks, are typically not considered a threat against fully synthetic data since fully synthetic records do not correspond to real individuals (Hu, 2019). However, attribute disclosure attacks can still leverage synthetic data to deduce information about targeted individuals and this type of attack is commonly considered in the literature (Quick et al., 2015; Hittmeir et al., 2020; Taub et al., 2018; Guo and Hu, 2022). We wish to assess whether synthetic data that is legally anonymous per the criteria described above also provides protection against attribute disclosure.

We define an attribute disclosure attack as follows. Suppose an adversary is interested in deducing the value of some sensitive categorical or intervaled attribute S_j for a record of interest $\mathbf{y}_j$. We denote the true value of S_j as s_j^*. The adversary observes the synthetic records in $\mathbf{Z}$ but does not observe the confidential record $\mathbf{y}_j$. From the perspective of the adversary, S_j is a random variable which takes on any value s_j from the sample space $\mathcal{S}$. The adversary has some prior probability mass function over the values of S_j, denoted $p(S_j|\mathbf{x}_j^q, \mathbf{b})$, where $\mathbf{x}_j^q$ denotes the quasi-identifying variable values available to the adversary that are also contained in the confidential record $\mathbf{y}_j$. We let $\mathbf{b}$ denote any other background information available to the adversary. Upon observing a synthetic data set $\mathbf{Z}$, the adversary predicts the value of S_j using $p(S_j|\mathbf{Z}, \mathbf{x}_j^q, \mathbf{b})$, the probability mass function for S_j conditional on the external information, background information, and the synthetic data. It is typically assumed that the adversary predicts the value of S_j to be the value s_j' with the highest probability,

$$s_j' = \text{argmax}_{s_j} p(S_j = s_j \mid \mathbf{Z}, \mathbf{x}_j^q, \mathbf{b}). \tag{13}$$

From this, a measure of attribute disclosure risk is the percentage of records for which the adversary correctly infers the value s_j^*,

$$R = \frac{\sum_{j=1}^{N_1} I(s_j' = s_j^*)}{N_1}. \tag{14}$$

However, this is a problematic measure if some values of a sensitive attribute occur significantly more often than others, *e.g.*, for a binary variable that is equal to 0 in most cases, an adversary could make a high percentage of accurate predictions by inferring the value 0 for every confidential record without observing $\mathbf{Z}$.

Instead, we wish to bound the *increase* in probability of inferring the correct value s_j^* for each sensitive record, conditional on observing $\mathbf{Z}$. The probability of such an attack was bounded via differencing in Bale et al. (2022), which can allow for large increases in the relative probability of inferring the correct value s_j^*. Instead, as suggested by (Reiter et al., 2014), we measure the multiplicative increase in the probability of inferring s_j^* in legally anonymized synthetic data, and assume that the maximum increase over all N_1 confidential records should be bounded as follows,

$$\max \quad \{j = 1, \ldots, N_1 \,:\, p(S_j = s_j^*|\mathbf{Z}, \mathbf{x}_j^q, \mathbf{b}) / p(S_j = s_j^*|\mathbf{x}_j^q, \mathbf{b})\} \leq c. \tag{15}$$

The value of c should be greater than or equal to 1, and the strength of privacy protection decreases as c increases.

The prior probability $p(S_j = s_j^*|\mathbf{x}_j^q, \mathbf{b})$ should be chosen to serve as a reasonable baseline for the ability of an adversary to infer values of the sensitive variable S_j without access to the synthetic data set (Bale et al., 2022). Examples

could be random guessing with equal probability over the possible values of S_j, or using publicly available population level estimates of the probabilities of the possible values S_j. To calculate $p(S_j = s_j^* | \mathbf{Z}, \mathbf{x}_j^q, \mathbf{b})$, we assume the adversary constructs the following δ-neighborhood,

$$\mathcal{N}(\mathbf{x}_j^q; \delta_{AD}, \mathbf{Z}) = \{i = 1, \ldots, N_2 \, : \, d(\mathbf{x}_j^{q(cont)}, \mathbf{z}_i^{q(cont)}) \leq \delta_{AD} \wedge I(\mathbf{x}_j^{q(cat)} = \mathbf{z}_i^{q(cat)}) = 1\}. \tag{16}$$

The distance measure $d(\mathbf{x}_j^{q(cont)}, \mathbf{z}_i^{q(cont)})$ (*e.g.*, Euclidean distance) computes the similarity between the continuous quasi-identifying attributes $\mathbf{x}_j^{q(cont)}$ and $\mathbf{z}_i^{q(cont)}$ of the external information and ith synthetic record, respectively. The expression $I(\mathbf{x}_j^{q(cat)} = \mathbf{z}_i^{q(cat)})$ is equal to one when the categorical and discrete quasi-identifying attributes of the external information and ith synthetic record are identical. We assume the adversary computes the probability mass function $p(S_j = s_j^* | \mathbf{Z}, \mathbf{x}_j^q, \mathbf{b})$ as the probability distribution of values of S_j from the synthetic records which are contained in $\mathcal{N}(\mathbf{x}_j; \delta_{AD}, \mathbf{Z})$.

In order to meet (15), we seek to bound the multiplicative increase in the probability of the adversary correctly inferring that $S_j = s_j^*$. We define an estimate of this probability, conditional on $\mathbf{x}_j^q$, $\mathbf{b}$, $\mathbf{Z}$, and a closeness threshold δ_{AD} as follows,

$$p(S_j = s_j^* | \mathbf{Z}, \mathbf{x}_j^q, \mathbf{b}, \delta_{AD}) = \frac{\sum_{\mathbf{z}_i \in \mathcal{N}(\mathbf{x}_j^q; \delta_{AD}, \mathbf{Z})} I(s_i = s_j^*)}{|\mathcal{N}(\mathbf{x}_j^q; \delta_{AD}, \mathbf{Z})|}. \tag{17}$$

where $I(.)$ is the indicator function. This estimate is equivalent to the proportion of synthetic records in the δ-neighborhood of $\mathbf{x}_j^q$ with sensitive attribute values matching s_j^*. If there are no synthetic records in the $\delta-$neighborhood, the adversary either can not perform the attack or increases the value of δ_{AD}. We note that this estimate is a generalization of the "Type S" risk assessed by Quick et al. (2015). The choice of distance measure $d(.)$ and closeness threshold δ_{AD} must be selected by the data controller. The δ_{AD} parameter should be selected to provide a reasonable balance between data utility and privacy, and should be adjusted based on the dimensionality of the data.

4 | SIMULATION STUDY

We perform a simulation study similar to that of Snoke et al. (2018) to verify that the privacy criteria in equations (5), (6), and (8) are met when $\mathbf{Y}$, $\mathbf{H}$, and $\mathbf{Z}$ are sampled independently from the same data generating distribution $\mathcal{D}$. We perform 1,000 simulations in total. For each simulation, we sample ten confidential, ten synthetic, and ten holdout data sets containing 5,000 records each from a ten-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu} = 0$, variances of $\sigma^2 = 1$, and covariances of the ith confidential/synthetic/holdout data set being equal to $0, 0.1, \ldots, 0.9$ for $i = 1, \ldots, 10$. We then calculate the identical match share using $\delta_{IMS} = 1$, the fifth percentile of the DCR, and the fifth percentile of the NNDR for the ith confidential, synthetic, and holdout data sets for $i = 1, \ldots, 10$. Table 1 contains the difference in the averages of each privacy criterion across all synthetic and holdout data sets.

We use the pMSE ratio to verify the distributional similarity of the sampled data sets. However, we use the expected pMSE derived by Snoke et al. (2018) in their Appendix A.2 (the second case mentioned in Section 2), which is the expected pMSE between data sets sampled from the same data generating distribution. This value is equivalent to twice the value shown in Equation (2),

$$\mathbb{E}(pMSE) = 2 * (k-1)(1-c)^2 c/N, \tag{18}$$

and amounts to treating both data sets as matrices of random variables. We use this value of the expected pMSE because it identifies when two data sets have been directly sampled from the same data generating distribution. In other words, the expected pMSE in 2 assumes the confidential data $\mathbf{Y}$ was generated from some data generating distribution $f(\mathbf{Y}; \boldsymbol{\theta}) = \mathcal{D}$, and the synthetic data is sampled from a synthesis model $f(\mathbf{Y}; \hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is estimated from $\mathbf{Y}$. Instead, we seek to treat the confidential data and synthetic data as two samples from the same synthesis model, such that the criteria in Section 3 are met. We compute the pMSE ratio using the expected pMSE from Equation (18) between the confidential and synthetic data sets from each simulation and covariance value using a logistic regression with three way polynomial terms and interactions. We compute the average ratio across all simulations for each covariance value and report the results in Table 1.

Covariance	IMS	DCR	NNDR	pMSE Ratio
0.0	0.0000	0.0003	0.0002	1.0057
0.1	-0.0001	0.0005	0.0001	1.0068
0.2	0.0000	0.0001	0.0002	1.0061
0.3	0.0000	0.0001	-0.0002	1.0065
0.4	-0.0004	0.0004	0.0002	1.0095
0.5	0.0001	0.0000	0.0002	1.0072
0.6	-0.0003	0.0000	-0.0004	1.0060
0.7	0.0002	-0.0004	-0.0005	1.0130
0.8	0.0001	0.0000	0.0002	1.0110
0.9	0.0000	0.0000	-0.0001	0.9946

TABLE 1 Results from simulation study. Difference (synthetic average - holdout average) of mean privacy measures. The pMSE ratios are calculated based on the synthetic data relative to the training data using a logistic regression with three way polynomial terms and interactions, and average across all simulations.

The simulation results in Table 1 confirm the analytical results from Section 3. When the synthetic data is drawn from the same data generating distribution as the holdout data (and the confidential data) then the expected values of the identical match share and DCR and NNDR percentiles are approximately equal between the synthetic and holdout data sets. We also confirmed the results from Snoke et al. (2018) that the ratio of the pMSE between synthetic data and confidential data generated from the same distribution, and the expected pMSE from Equation (18), is approximately one.

5 | SYNTHESIS METHOD FOR LEGALLY ANONYMOUS SYNTHETIC DATA

5.1 | Model for Continuous Variables

We select a Gaussian mixture model (GMM) for synthesizing continuous variables since it can accurately estimate the density of multi-modal distributions. To our knowledge, the only other application of GMMs for data synthesis was performed by (Oganian and Domingo-Ferrer, 2017), who use a GMM to generate synthetic data that satisfies probabilistic k-anonymity.

Assume the confidential data set $\mathbf{Y}$ contains a subset of P continuous attributes denoted $\mathbf{Y}^{cont} = [Y_1, \ldots, Y_P]$. Let $\mathbf{y}_j^{cont} = (y_{j,1}, \ldots, y_{j,P})$ denote the continuous attribute values contained in the confidential record $\mathbf{y}_j$. A GMM with K mixture components approximates the joint probability density function $p(\mathbf{y}^{cont})$ as:

$$p(\mathbf{y}^{cont}) = \sum_{k=1}^{K} \pi_k N(\mathbf{y}^{cont}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{19}$$

Here, $N(\mathbf{y}^{cont}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a multivariate Gaussian distribution (the kth mixture component), π_k is the weight of the kth mixture component, and $\sum_{k=1}^{K} \pi_k = 1$ so that $p(\mathbf{y}^{cont})$ is a proper probability density. The data owner must choose a value of K prior to estimating the model; higher values of K correspond to more mixture components and can lead to overfitting, whereas lower values of K may produce a biased model that poorly approximates a multimodal data distribution.

To estimate the model, each record $\mathbf{y}_j^{cont}$ is assumed to be i.i.d. and we obtain the parameters $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^{K}$ that maximize the log-likelihood of observing $\mathbf{Y}^{cont}$ by solving the following problem using expectation maximization (EM)[3]:

$$\begin{aligned} \max_{\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^{K}} \quad & \sum_{j=1}^{N_1} \log \sum_{k=1}^{K} \pi_k N(\mathbf{y}_j^{cont}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \text{s.t.} \quad & \sum_{k=1}^{K} \pi_k = 1 \\ & \pi_k \geq 0, \quad \forall k. \end{aligned} \tag{20}$$

Next, we sample $\mathbf{m} = (m_1, \ldots, m_K)$ from a multinomial distribution $M(\mathbf{m}; N_2, \boldsymbol{\pi}^*)$, and then for each $k = 1, \ldots, K$, we sample m_k synthetic records from $N(\mathbf{y}^{cont}; \boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*)$ to create $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{N_2})$.

5.2 | Models for Categorical and Discrete Variables

We consider both multinomial logit (MNL) models and classification and regression trees (CART) for synthesizing categorical and discrete variables. We assume the confidential data set $\mathbf{Y}$ contains a subset of Q categorical and/or discrete attributes denoted $\mathbf{Y}^{cat} = [Y_{P+1}, \ldots, Y_{P+Q}]$. Let $\mathbf{y}_j^{cat} = (y_{j,P+1}, \ldots, y_{j,P+Q})$ denote the categorical/discrete attribute values contained in the confidential record $\mathbf{y}_j$. The MNL and CART synthesizers require a sequential synthesis approach (one variable at a time) which differs from the joint synthesis approach described for GMMs. The sequential synthesis approach is equivalent to approximating the original data distribution using the product of conditional

[3] See the appendix for an overview of the EM algorithm for GMMs

univariate distributions (Jackson et al., 2022):

$$p(Y_{P+1}, \ldots, Y_{P+Q}) = p(Y_{P+1}) \prod_{q=P+2}^{P+Q} p(Y_q | Y_{q-1}, \ldots, Y_{P+2}, Y_{P+1}). \tag{21}$$

To implement the sequential approach, a synthetic Z_{P+1} is typically sampled from the marginal distribution of Y_{P+1}. The second variable Y_{P+2} is modeled conditional on Y_{P+1} using a MNL or CART, and the synthetic Z_{P+2} is sampled from the synthesis model conditional on Z_{P+1}. Next, Y_{P+3} is modeled conditional on Y_{P+2} and Y_{P+1}, Z_{P+3} is sampled conditional on Z_{P+2} and Z_{P+1}, and so on.

5.2.1 | Multinomial Logit (MNL) Model

Suppose we are synthesizing categorical variable Y_q conditional on the variables $\mathbf{Y}'^{cat} = [Y_{q-1}, \ldots, Y_{P+2}, Y_{P+1}]$. The objective function for the MNL model is

$$\min_{\{\beta_k\}_{k=1}^K} \quad -C \sum_{j=1}^{N_1} \sum_{k=1}^{K} I(y_{j,q} = k) \log(\hat{p}(y_{j,q} = k \mid \mathbf{y}_j'^{cat})) + \sum_{k=1}^{K} ||\beta_k||_1, \tag{22}$$

where β_k is the vector of coefficients for category k and C is the inverse of the regularization strength. Other implementations of this synthesizer replace the parameter C with a parameter λ which directly multiplies the penalty $\sum_{k=1}^{K} ||\beta_k||_1$ (Schneider et al., 2023). The probability $\hat{p}(y_{j,q} = k \mid \mathbf{y}_j'^{cat})$ that the value of variable Y_q is equal to k for record j is predicted as:

$$\hat{p}(y_{j,q} = k \mid \mathbf{y}_j'^{cat}) = \frac{\exp(\mathbf{y}_j'^{cat} \beta_k)}{\sum_{k=1}^{K} \exp(\mathbf{y}_j'^{cat} \beta_k)}. \tag{23}$$

To generate a synthetic version of Y_q, denoted Z_q, a synthetic value for each record j is sampled from a multinomial distribution defined by the probabilities $\{k = 1, \ldots, K : \hat{p}(y_{j,q} = k \mid \mathbf{y}_j'^{cat})\}$.

The parameter C is used to control the tradeoff between privacy and utility. Large values of C heavily weight the contribution of the log probabilities to the objective (22), producing an overfitted model with high utility that could reproduce the original variable Y_q. Small values of C place a heavier weight on the magnitude of the coefficients; this can produce a biased model where all coefficients are forced to zero and synthesis becomes equivalent to randomly sampling from the marginal distribution of Y_q independent of $\mathbf{Y}'^{cat}$.

5.2.2 | Classification and Regression Tree (CART) Models

The description in this section borrows heavily from that of (Drechsler and Reiter, 2011; Drechsler and Hu, 2021). The CART synthesizer approximates the distribution of Y_q conditional on $\mathbf{Y}'^{cat}$ by partitioning the predictor space into subsets of records with values of Y_q that are as homogenous as possible. The tree is grown using recursive binary splits on the predictors in $\mathbf{Y}'^{cat}$ that successively minimize the Gini index; splitting stops when the deviance in all leaves is less than some threshold or when there can not be at least b records in each child leaf. To generate Z_q, the

values of Y_q in each leaf are resampled using a Bayesian bootstrap (Rubin, 1981; Reiter, 2005) and a synthetic value for each record is randomly drawn from the bootstrapped values in the appropriate leaf.

A data controller can leverage many different parameters, such as the maximum tree depth, the minimum number of records required to perform a split, the minimum number of records b required in each leaf node (Reiter, 2005), and the complexity parameter cp (Drechsler and Hu, 2021) to control the tradeoff between privacy and utility. Any of these parameters could be used to control the size, or depth, of a CART, where deeper trees will produce synthetic data with higher utility and lower privacy, with the complexity parameter cp being used in recent work (Drechsler and Hu, 2021; Schneider et al., 2023).

5.3 | Synthesis for Mixed Data Sets

For data sets with both continuous and categorical/discrete variables, we combine the GMM and MNL/CART based approaches. The continuous variables are first jointly synthesized using the GMM, and the remaining categorical/discrete variables can be modeled using a sequential approach where the synthesis is conditional on the synthesized continuous variables, *i.e.*, the synthesized continuous variables take the place of Y_{P+1} in Equation 21. This is the approach we use in our empirical application in Section 6.

5.4 | Bayesian Optimization of Synthesis Models

We propose using Bayesian optimization based on a Gaussian process (Snoek et al., 2012) to optimize the privacy parameters of our proposed synthesis method to correctly approximate the data generating distribution $\mathcal{D}$. Bayesian optimization is a popular hyper-parameter optimization method for black-box machine learning models. In general, it is used to optimize an objective function without a closed form expression, but from which we can observe the objective function at sampled values. A thorough review of Bayesian optimization is given by Brochu et al. (2010).

Our analysis in Section 3 showed that the legal privacy criteria are met in expectation when the synthetic data is sampled from the data generating distribution $\mathcal{D}$. Therefore, we will use the Bayesian optimization process to choose parameters that yield synthesis models that approximate $\mathcal{D}$. An objective function $\boldsymbol{f}$ is used to calculate the distributional similarity between the synthesized data and the original data to determine whether we have a good approximation of $\mathcal{D}$.

We define $\boldsymbol{\theta}_i$ as the ith sample of synthesis model parameters, where the objective $\boldsymbol{f}(\boldsymbol{\theta}_i) = f_i$ gives the measure of distributional similarity at $\boldsymbol{\theta}_i$. It is assumed that $\boldsymbol{f}(\boldsymbol{\theta})$ is drawn from a Gaussian process prior. The Bayesian optimization method estimates the objective function using the posterior mean function of a Gaussian process. The optimization is carried out by sequentially choosing values of $\boldsymbol{\theta}$ which maximize the expected improvement in the objective function $\boldsymbol{f}(\boldsymbol{\theta}_i)$. We define the objective function $\boldsymbol{f}$ using the pMSE ratio as follows:

$$\min_{\{\boldsymbol{\theta}\}} \quad (1 - \frac{1}{m}\sum_{i=1}^{m} pMSE_m / \mathbb{E}[pMSE])^2, \tag{24}$$

which minimizes the squared deviation from one of the average pMSE ratio across m synthetic data sets from one, where $\boldsymbol{\theta}$ denotes the parameters for the synthesis model(s). Note that we utilize the expected pMSE between two data sets directly sampled from the same data generating distribution described in Section 4.

Optimizing a synthesis method using Bayesian optimization is similar conceptually to training a generative adversarial network, (*e.g.*, Anand and Lee (2023). GANs are a complex data synthesis method and do not have a single pa-

rameter that data controllers can use to leverage the privacy/utility trade-off. Instead, the GAN is trained to minimize the probability of the discriminator network incorrectly classifying a synthetic observation as real, and to maximize the probability that the generator network produces synthetic observations that will be labeled as real. Using this method, in theory, the distribution of the synthetic data will converge to the true data distribution (Goodfellow et al., 2020). Likewise, we seek to optimize our synthesis approach such that it generates synthetic data that is indistinguishable from another data set that was directly sampled from the same data generating distribution as the confidential data.

5.5 | Inference Prevention Algorithm

We leverage the ability to sample synthetic locations from individual mixture components in the GMM synthesizer to create an algorithm that reduces the probability of successful attribute disclosure. For now, we implement an algorithm that is applicable to preventing attribute disclosure against a binary sensitive attribute. The algorithm works by generating synthetic records within the neighborhood $\mathcal{N}(\mathbf{x}_j^q; \delta_{AD}, \mathbf{Z})$ that have different sensitive attribute values than the confidential record $\mathbf{y}_j$, thereby reducing the probability ratio in (15). To generate synthetic records in the neighborhood $\mathcal{N}(\mathbf{x}_j^q; \delta_{AD}, \mathbf{Z})$, we find the mixture component with the highest responsibility for record $\mathbf{y}_j^{cont}$:

$$r_{j,k} = \frac{\pi_k \mathcal{N}(\mathbf{y}_j^{cont}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(\mathbf{y}_j^{cont}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}, \tag{25}$$

which is the probability that $\mathbf{y}_j^{cont}$ was generated by the kth mixture component. We then sample a large number of candidate synthetic records (SS in Algorithm 1) and keep the minimum number of candidate records (ss in Algorithm 1) which are the closest to $\mathbf{y}_j^{cont}$ required to meet the criterion in (15). We assign these candidate points the same values $\mathbf{y}_j^{cat}$, and assign sensitive attribute values $S = s_j^* - 1$.

6 | EMPIRICAL APPLICATION

6.1 | The Dataset: South Korean COVID-19 Patient Location Data

During the COVID-19 pandemic, the South Korean government released sensitive data on COVID-19 patients in an attempt to mitigate disease spread (Zastrow, 2020). These releases came in the form of alerts sent to individuals living near an infected person, detailing the infected person's age, gender, and movement histories. Such detailed data releases may have helped increase the public's trust in the South Korean government and enabled individuals to know if they had been in contact with an infected individual. COVID-19 mapping applications, such as the one pictured in Figure 1, were built on this data and helped the public know where they may have been exposed to COVID-19.

Algorithm 1: Inference Prevention Algorithm

Data: $\mathbf{Y}$, $\mathbf{Z}$, $p(\mathbf{y}^{cont})$, δ_{AD}, $p(S_j = s_j^* | \mathbf{x}_j^q, \mathbf{b})$ for all possible s_j, c, SS

for $\mathbf{y}_j \in \mathbf{Y}$ **do**

- $nn \leftarrow \mathcal{N}(\mathbf{x}_j^q; \delta_{AD}, \mathbf{Z})$; // calculate the δ-neighborhood of $\mathbf{y}_j$
- $nnum \leftarrow |\mathcal{N}(\mathbf{x}_j^q; \delta_{AD}, \mathbf{Z})|$; // calculate the number of synthetic neighbors
- $prob \leftarrow p(S_j = s_j^* | \mathbf{Z}, \mathbf{x}_j^q, \mathbf{b}) / p(S_j = s_j^* | \mathbf{x}_j^q, \mathbf{b})$; // calculate ratio from Equation (15) for $\mathbf{y}_j$
- **if** $prob > c$ **then**
  - $m \leftarrow p(S_j = s_j^* | \mathbf{Z}, \mathbf{x}_j^q, \mathbf{b}) \times nnum$; // number of neighbor synthetic records with s_j^*
  - $ss \leftarrow \lceil m/(c \times p(S_j = s_j^* | \mathbf{x}_j^q, \mathbf{b})) - nnum \rceil$; // neighbors needed with $s \neq s_j^*$ to reach c
  - $\mathbf{Z}^{new(cont)} \leftarrow \{\mathbf{z}_1^{new(cont)}, \ldots, \mathbf{z}_{ss}^{new(cont)}\}$; // sample $SS >> ss$ records from $p(\mathbf{y}^{cont})$, // keep the ss records closest to $\mathbf{y}_j^{cont}$
  - $\mathbf{Z}_{new} \leftarrow \{i = 1, \ldots, ss : (\mathbf{z}_i^{new(cont)}, \mathbf{y}_j^{cat})\}$; // append $\mathbf{y}_j^{cat}$ to new records
  - **for** $i = 1, \ldots, ss$ **do**
    - $S_i^{new} \leftarrow 1 - s_j^*$; // set sensitive attribute for new records to $s \neq s_j^*$
  - **end**
  - $\mathbf{Z} \leftarrow [\mathbf{Z}, \mathbf{Z}_{new}]$; // append $\mathbf{Z}_{new}$ to $\mathbf{Z}$
- **end**

end

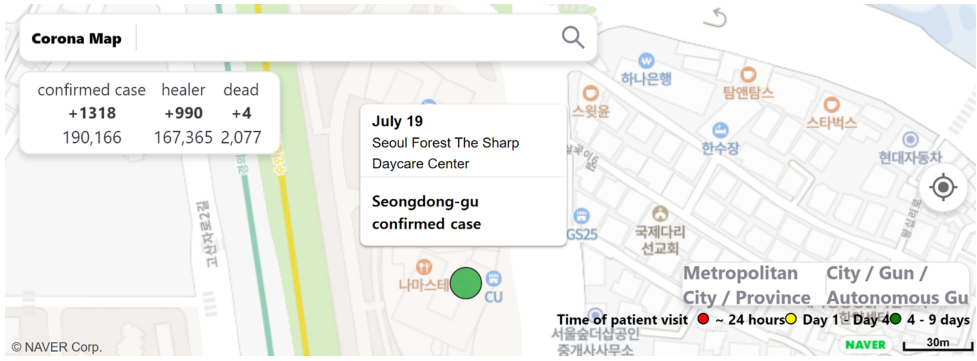

FIGURE 1 Screenshot from coronamap.site.

Unfortunately, even this well-intentioned release of sensitive data harmed data subjects. Feelings of privacy invasions dissuaded individuals from getting tested, and non-infected individuals avoided places that an infected person was known to have visited long after it was safe to return (Zastrow, 2020). There were also severe consequences for data subjects; accusations of infidelity, insurance fraud, and prostitution arose from the data alerts (Kim, 2020).

Shortly after the start of the COVID-19 pandemic, a group of researchers released a number of datasets containing information on the COVID-19 pandemic in South Korea[4]. The effort was labeled "Data Science for COVID-19 (DS4C)". The data sets included population numbers, search trends, regional weather information, time series of overall COVID-

[4]The Kaggle page containing the data can be found here.

19 patient status (whether a patient is surviving or deceased) by province, gender, and age, and patient-level location and demographic information. The goal of the data release was presumably to enable widespread analysis of the data in an attempt to gain helpful insights early in the pandemic. Various data analysis competitions were held and several research projects made use of the data (e.g., (Husnayain et al., 2020)). Some time after the data release, however, it was determined that the patient-level location data represented a privacy risk and was removed from the release.

While the data released as part of the DS4C effort was highly sensitive, it was also highly useful for research. We obtained a copy of the full data, including the patient-level location data, and our goal is to use our optimized sequential synthesis strategy to generate a synthetic version of the data which could be released in place of the sensitive data. The sensitive data set is comprised of *location measurements* (latitude/longitude), *sex* (Male/Female), *age* (ranging from 0 to 99 in buckets of 10 values), and the *state* (whether they passed away or recovered from the disease) of COVID-19 patients in South Korea. In total, we have $N = 6,712$ records, 46.2% of which correspond to males, and *55* of which correspond to deceased individuals. The distribution of *age* is shown in Figure 2, where most values fall in the range of 20s to 60s.

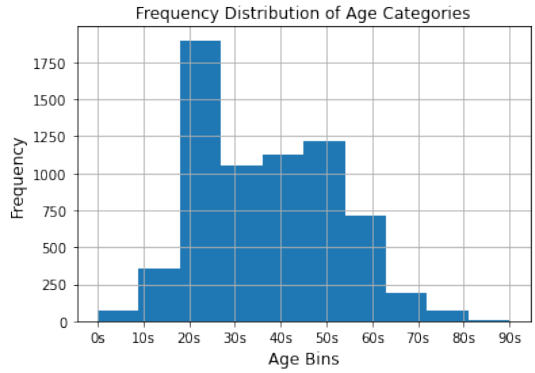

FIGURE 2 Frequency distribution of age variable.

For the purposes of this paper, we simulate sampling data sets from the same data generating distribution by sampling (with replacement) $N = 3,361$ records (the training data set $\mathbf{T}$) from the original data. We create $M = 20$ holdout data sets $\mathbf{H}_m, m = 1, \ldots, 20$, of the same size using the same sampling strategy for comparison with the synthetic data sets $\mathbf{Z}_m, m = 1, \ldots, 20$ from each of the synthesis methods.

6.2 | Synthesizer Implementation

We use the sequential synthesis approach outlined in Section 5. All models were implemented using scikit-learn (Pedregosa et al., 2011). A GMM is used to synthesize the continuous location variables (latitude, longitude). The mixture model is initialized five different times using k-means++ (Arthur and Vassilvitskii, 2007) and is optimized using the expectation maximization algorithm.[5]

We compare two different approaches to synthesizing the remaining variables (sex, age, state). The first approach uses MNLs as described in Section 5 using three-way interactions and polynomial terms, where the privacy-utility tradeoff is controlled by the inverse regularization parameter C. The second approach uses CARTs with no limits on

[5] See the appendix for an outline of the expectation maximization algorithm applied to GMMs.

the depth or the *cp* parameter. In our preliminary analysis, we found the trade-off between privacy and utility to be much more responsive to adjustments to b than adjustments to cp, which is consistent with Drechsler and Hu (2021) who found cp to be ineffective at tuning the balance between privacy and utility. Rather, we control the privacy-utility trade-off using b, the minimum number of records required in any leaf node.

To implement the Bayesian optimization method, we use the `bayesian-optimization` python package (Nogueira, 2014). The parameter ξ, which is used to balance between exploration (testing $\boldsymbol{\theta}$ where $\boldsymbol{f}$ has large uncertainty) and exploitation (testing $\boldsymbol{\theta}$ where $\boldsymbol{f}$ is expected to be high), is set to 0.01, as recommended in Brochu et al. (2010). We perform the optimization process for five random seeds. For each seed, the optimization process is initialized with five random points, and then performs 25 iterations. Based on the $\boldsymbol{\theta}$ in each iteration, we evaluate $\boldsymbol{f}$ over twenty synthetic datasets. The minimum objective value for the MNL and CART synthesis strategies across iterations for each random seed is plotted in Figure 3. We select the $\boldsymbol{\theta}^*$ with the best overall $\boldsymbol{f}$ value, shown for each synthesis strategy in Table 2.

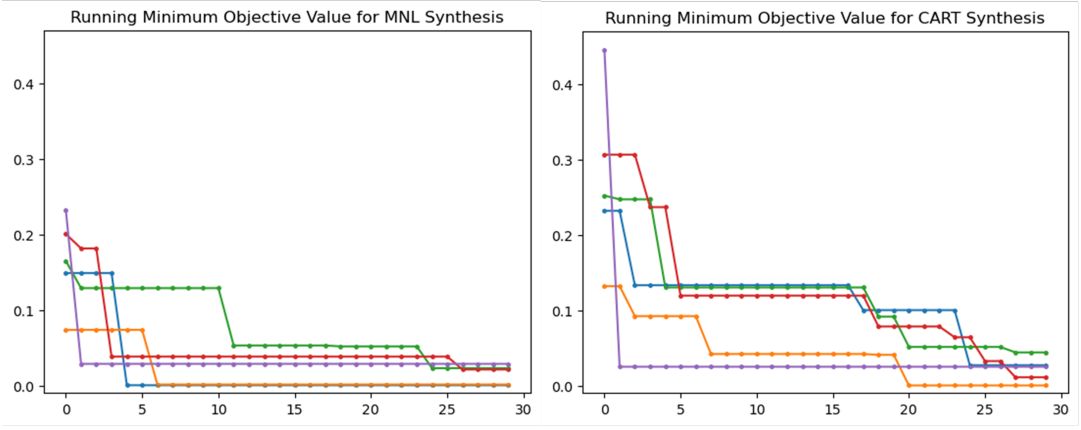

FIGURE 3 Running minimum of objective value for the MNL (left) and CART (right) synthesizers.

Both synthesis strategies were able to obtain objective values relatively close to zero, with the CART based strategy producing an average pMSE ratio marginally closer to one. Both synthesis strategies applied the most regularization in the models for Sex. The MNL based strategy applied to least regularization in the model for State, whereas the CART based strategy applied to least regularization to the model for Age. The synthesis strategies selected very different numbers of mixture components, but the effect appears to be minimal on the later privacy and utility results.

Variable	GMM + MNL	GMM + CART
Objective	$f = 0.0003$	$f = 0.0001$
Lat/Long	$K = 730$	$K = 524$
Sex	$C = 0.029$	$b = 158$
Age	$C = 1.007$	$b = 7$
State	$C = 1.546$	$b = 15$

TABLE 2 Objective values and parameters selected by the Bayesian optimization process for the multinomial based synthesis process (GMM + MNL) and the CART based synthesis process (GMM + CART).

7 | SYNTHESIS RESULTS

We synthesize twenty synthetic data sets using each synthesis strategy, including the MOSTLY.AI method, and compare the synthetic data sets on privacy and utility metrics. For the calculation of the privacy metrics, we use the euclidean distance between the synthetic and confidential and holdout and confidential data sets. Prior to calculating the distances, we standardize the confidential data set, and standardize the synthetic and holdout data sets using the means and standard deviations from the confidential data.

7.1 | Privacy Criteria

7.1.1 | Legal Privacy Criteria

Figure 4 plots the average IMS across the twenty synthetic data sets from each synthesis method for a range of δ_{IMS} values. The synthesis methods denoted – `Inference` indicate that the Inference prevention algorithm was applied to the corresponding synthetic data sets. The average IMS for all synthesis methods is lower than or equal to the average IMS for the holdout data sets for all δ_{IMS} values. Therefore, all of the synthetic data sets under consideration meet the IMS criterion (5) on average. The IMS is noticably higher for the CART based synthesis strategy than the other synthesis strategies for smaller values of δ_{IMS}. Applying the Inference prevention algorithm raises the average IMS for the CART synthesized data sets, and decreases it for the MNL synthesized data sets.

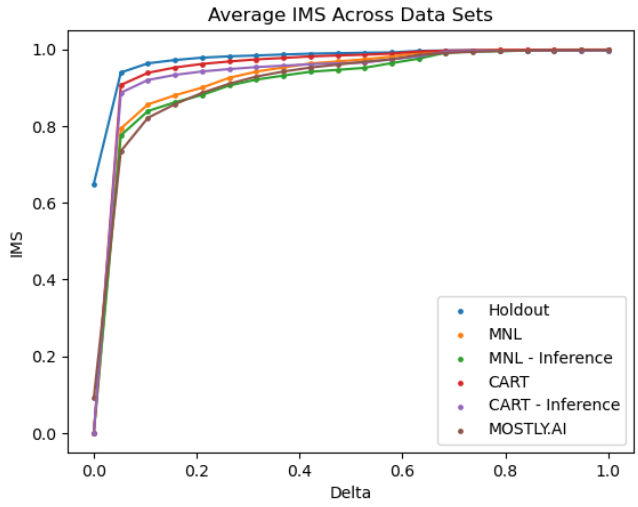

FIGURE 4 Average IMS value across $M = 20$ holdout data sets and synthetic data sets from each synthesis approach.

Table 3 contains the differences in the average fifth percentiles of the DCR and NNDR between the synthetic and holdout data sets for each synthesis method. The differences in the fifth percentile of the DCR distribution are small, with the MNL based strategy producing the largest differences between the synthetic and holdout data sets, followed by the CART and MOSTLY.AI based strategies. The effect of the Inference prevention algorithm on these values was negligible. None of the DCR distributions for the synthetic data sets (as measured by the fifth percentile) are closer to zero than the holdout data sets, indicating that the synthesis methods do not produce synthetic records that are systematically closer to the confidential data than the holdout data is to the confidential data.

Synthesis Type	DCR (Synthetic - Holdout)	NNDR (Synthetic - Holdout)
MNL Based	0.0006	0.0145
MNL Based - Inference	0.0006	0.0137
CART Based	0.0005	0.0081
CART Based - Inference	0.0005	0.0079
MOSTLY.AI	0.0000	0.0265

TABLE 3 Privacy results.

The results for the NNDR are qualitatively similar. The fifth percentile of the NNDR distribution for the holdout data is smaller than for the NNDR distributions for the synthetic data from all strategies, *i.e.*, holdout points tend to be closer to their nearest neighbors relative to their fifth nearest neighbors compared to the synthetic data sets. Compared to the DCR results, the MOSTLY.AI strategy actually provides the best privacy as measured by the NNDR. The MOSTLY.AI synthetic records tend to be closer to their nearest confidential neighbors than the other synthetic records, but they also tend to be closer to their fifth nearest confidential neighbors, producing the largest difference

in the average fifth percentiles of the NNDR distribution. There is a slightly higher tendency for the CART and MNL based strategies to produce synthetic points closer to their nearest confidential neighbors relative to their fifth nearest confidential neighbors, relative to the MOSTLY.AI strategy.

The privacy results for the MNL and MOSTLY.AI based strategies are stronger than the CART based strategy, since the CART based strategy had the second lowest difference in the average DCR percentile, the lowest difference in the NNDR percentile, and much higher average IMS at low values of δ_{IMS}. Overall, the synthetic data sets from every synthesis method meet the IMS, DCR and NNDR based criteria in (5), (6) and (8) on average. Therefore, any of the methods under study could be used to produce legally anonymous synthetic data. Preference should be given to the synthesis method that provides the best data utility and offers protection against attribute disclosure.

7.1.2 | Attribute Disclosure

We assume an adversary attempts to perform attribute disclosure and deduce the `state` S of records contained in the confidential data using the synthetic data. Using the attack described in Section 3.5, we simulate the adversary linking the external information $\mathbf{x}_j^q$ to the synthetic records in the corresponding δ-neighborhood $\mathcal{N}(\mathbf{x}_j^q; \delta_{AD}, \mathbf{Z})$. These synthetic records match the target record on $\mathbf{x}_j^{q(cat)} = [\texttt{age}_j, \texttt{sex}_j]$ and are within a distance $d(\mathbf{x}_j^{q(cont)}, \mathbf{z}_i^{q(cont)})$ based on `latitude` and `longitude` using the euclidean distance for the distance measure $d(.)$.

Using publicly available data from the World Health Organization[6], we use the total number of confirmed COVID-19 cases (9,583) and total number of deaths (152) in South Korea as of March 23, 2020 to calculate the prior probability $p(S_j = 1|\mathbf{x}_j^q, \mathbf{b}) = 152/9583 \approx 0.016$ and $p(S_j = 0|\mathbf{x}_j^q, \mathbf{b}) = 1 - 0.016 = 0.984$. We measure the maximum multiplicative increase in the probability of inferring the correct `status` (left-hand side of equation (15)) for each record $\mathbf{y}_j$ for a range of δ_{AD} values and plot the average across all synthetic data sets from each synthesis strategy in Figure 5. The maximum possible increase in inference occurs when an adversary correctly predicts with 100% probability that a record $\mathbf{y}_j$ has a status $S_j = 1$, *i.e.*, $p(S_j = 1|\mathbf{Z}, \mathbf{x}_j^q, \mathbf{b})/p(S_j = 1|\mathbf{x}_j^q, \mathbf{b}) = 1.00/0.016 = 62.5$. We apply the Inference prevention algorithm recursively until the criterion in (15) in met for all records for a given value of c, and do so for a range of c values for all synthetic data sets from the MNL and CART based approaches. Overall, the algorithm added less than 300 additional synthetic records to each synthetic data set.

In comparison to the holdout data sets, each synthesis method (without the Inference prevention algorithm) provides a modest reduction to the increase in inference probability. There are noticeable differences in the protection against attribute disclosure across synthesis methods. The MOSTLY.AI synthesis method provides only a slight reduction in inference probability relative to the holdout data sets. In particular, there is an extremely high inference probability even for $\delta_{AD} \approx 0$. This indicates that the MOSTLY.AI strategy is producing synthetic records which are extremely similar to the confidential records with $S_j = 1$ allowing the adversary to have a very high probability of inference. This is more problematic if we assume that an adversary's confidence in their prediction scales with the inverse of δ_{AD}, *i.e.*, if synthetic records are found within the δ-neighborhood of the targeted record for a very small δ_{AD}, the adversary is more confident in their prediction. For the MNL and CART based strategies, the increase in probability is very small for $\delta_A D \approx 0$.

[6] https://covid19.who.int/region/wpro/country/kr

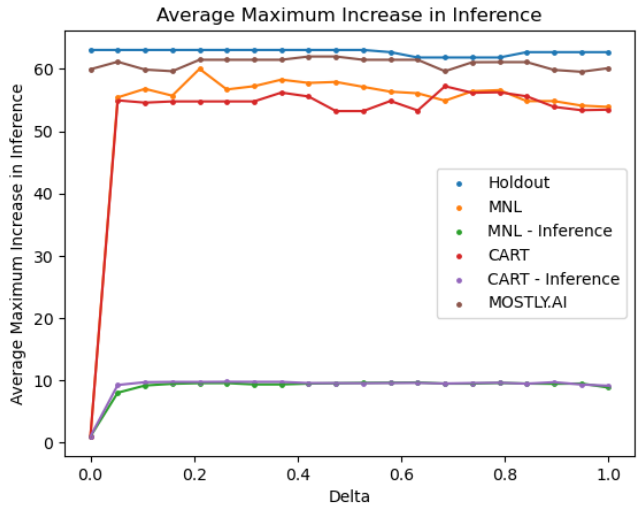

FIGURE 5 The average maximum multiplicative increase in inference probability across synthetic data sets and δ values for each synthesis method relative to the holdout data sets.

Figure 5 also shows that the Inference prevention algorithm works as intended: the average maximum increase in inference probability is bounded by $c = 10$ across all values of δ_{AD}. This is an extremely promising result since the it limits the probability $p(S_j = 1|\mathbf{Z}, \mathbf{x}_j^q, \mathbf{b})$ to only 0.16.

7.2 | Utility Measures

7.2.1 | Global Utility Measures

We assess the global utility of the synthetic data sets using two measures: (1) The pMSE ratio described in Section 5.4, which was used in the objective of the bayesian optimization process; and (2) the K and L functions utilized by Quick et al. (2015); Drechsler and Hu (2021) to assess the preservation of spatial properties in the synthetic data.

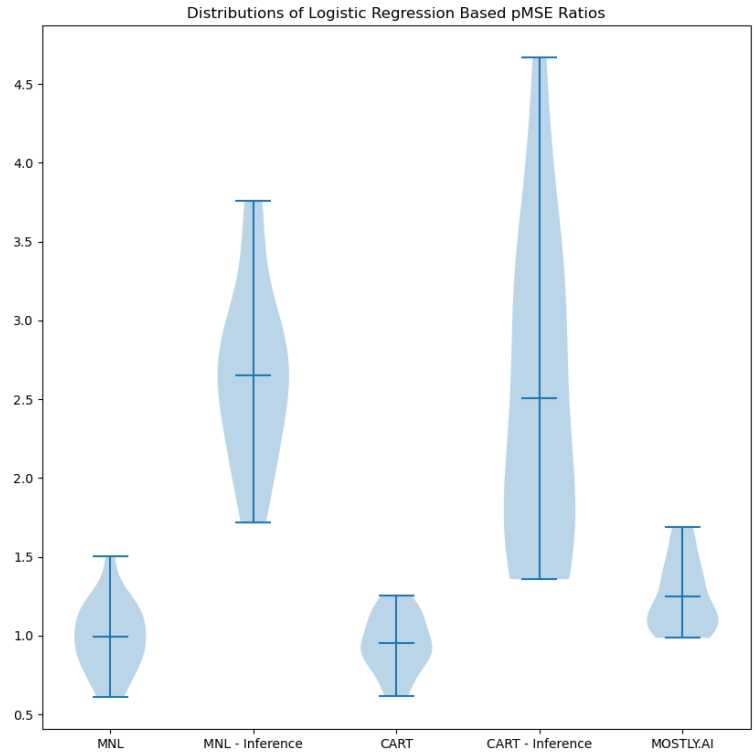

FIGURE 6 Distributions of pMSE ratios calculated using a CART model with null given by mean of pMSE under pairwise synthetic data sets.

For a baseline, we calculate the average pMSE ratio between the confidential and holdout data sets using the expected value of the pMSE under the null of data sets generated from the same synthesis model. This ratio comes out to 0.93, indicating a high distributional similarity between the holdout and confidential data sets.

Figure 6 shows the distributions of the pMSE ratios of each synthesis method relative to the confidential data set. The averages of the distributions under the MNL and CART based strategies are approximately one, which is consistent with the objective function values in Table 2. The MOSTLY.AI strategy has a slightly worse utility since the average pMSE ratio is approximately 1.25. The trade off between privacy and utility is evidenced by the large pMSE ratios that result from applying the Inference prevention algorithm, which average around 2.5. Overall, none of the synthesis methods produce synthetic data sets which are more similar to the confidential data than the holdout data is to the confidential data as measured by the propensity score ratio.

Next, we calculate the K and L functions. The K function is the expected number of points within a distance h of an arbitrary point. Letting ℓ_j denote the $(\texttt{latitude}, \texttt{longitude})$ tuple contained in record $\mathbf{y}_j$, an estimate of the K function can be obtained using the following:

$$\hat{K}(h) = \frac{|D|}{N} \sum_{j=1}^{N} \sum_{i=1}^{N} I(||\ell_j - \ell_i|| \leq h)/N, \text{ for } j \neq i. \tag{26}$$

Here, $I(.)$ is the indicator function which is equal to one whenever points ℓ_j and ℓ_i are within a distance h of each other, and $|D|$ is the area of the spatial domain. The L function can be estimated using the following:

$$\hat{L}(h) = \sqrt{\hat{K}(h)/\pi} - h, \tag{27}$$

where positive values of this estimate indicate spatial clustering. We estimate the values of the L function for the confidential data and each of the $m = 20$ synthetic data sets from each synthesis strategy for a range of h values. We include the mean and 95% empirical confidence interval estimates from the L function values calculated from the synthetic data sets in Figure 7. The left column contains results for all locations, and the right column contains results for the positive locations, *i.e.*, locations associated with records with $S = 1$ only. Changing the number of mixture components has little effect on the spatial clustering (comparing the the first and third plots of column 1). Applying the Inference prevention algorithm produces slight reductions in overall spatial clustering (comparing the first and second plots and third and fourth plots of the first column). The MOSTLY.AI synthesis strategy has much lower levels of spatial clustering than the locations synthesized using the GMM, indicating lower utility. The second column shows that the CART synthesis strategy provides the most accurate spatial clustering for the positive locations. Both the MNL and MOSTLY.AI strategies have lower levels of spatial clustering for the positive points.

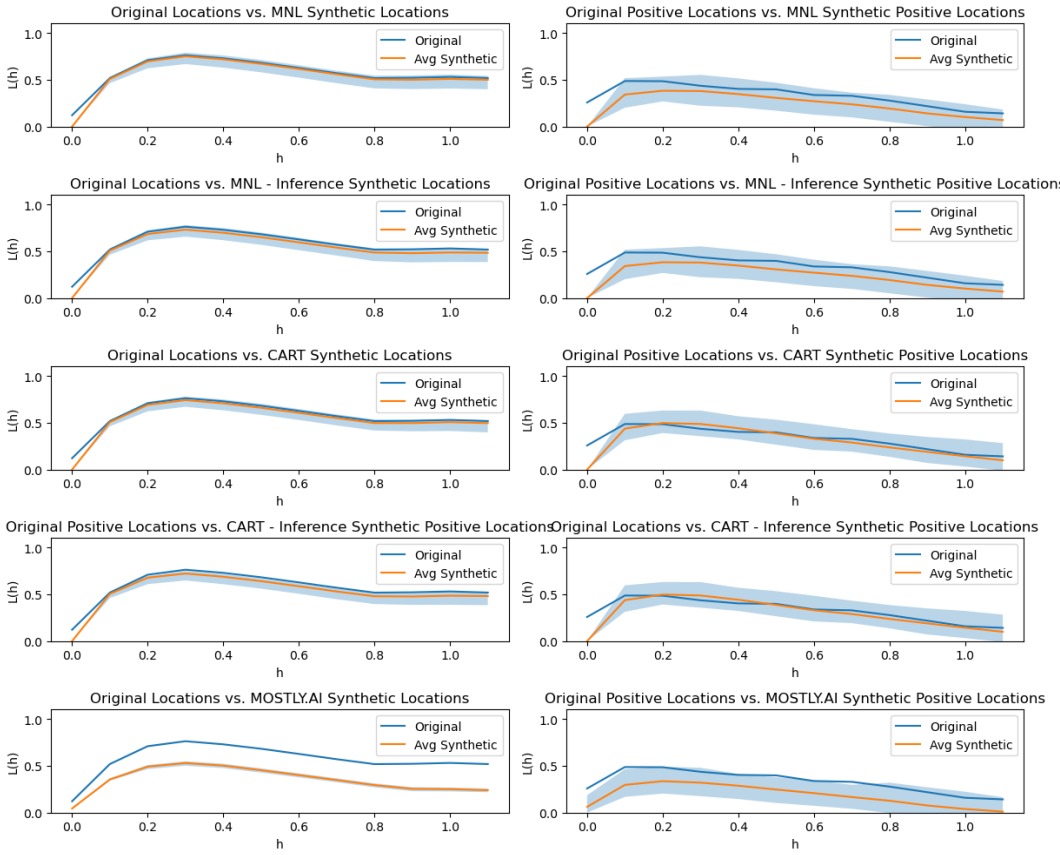

FIGURE 7 L-functions for the confidential data vs. average L-function values (orange) and 95% empirical confidence intervals (shaded) across synthetic data sets for each synthesis method.

The L-function results are confirmed when we visualize the confidential and synthetic locations. We selected one set of the synthetic locations from the GMM and one set from the MOSTLY.AI model for comparison to the confidential locations. The GMM synthesized locations (top right) bear a strong visual resemblance to the confidential locations (top left) and there is no visible changes after the Inference prevention algorithm is applied (bottom left), while the synthetic locations from the MOSTLY.AI synthesizer (bottom right) are more dispersed.

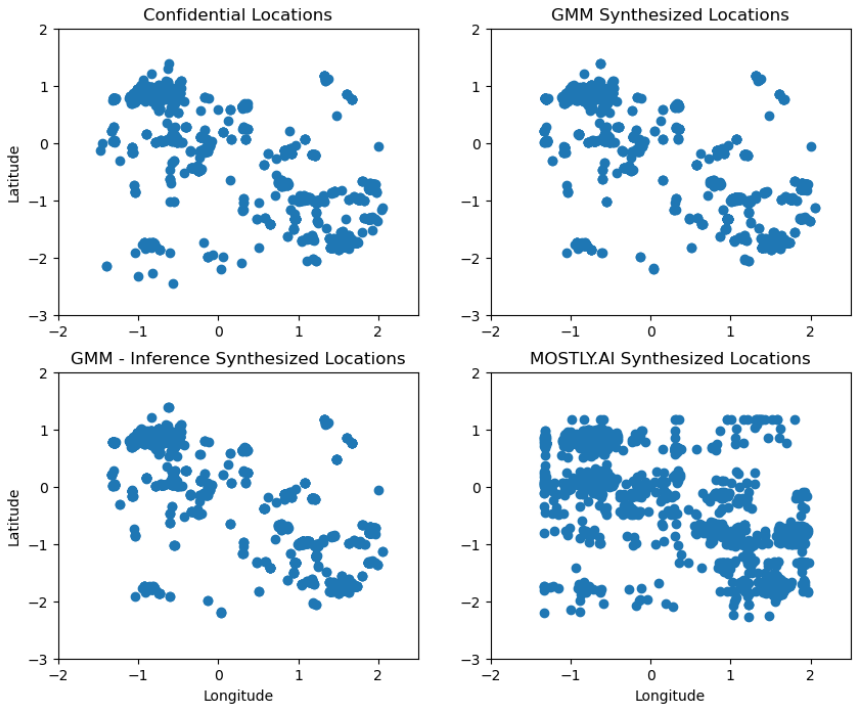

FIGURE 8 Comparison of confidential and synthetic locations from the GMM and MOSTLY.AI synthesis models.

7.2.2 | Analysis Specific Utility Measures

We evaluate analysis-specific utility by comparing the results of a data user's model applied to both the original and synthetic data sets. We utilize a simple logistic regression that predicts `status` as a function of `latitude`, `longitude`, `age`, `sex`, and iteractions between `latitude` and `longitude` and `sex` and `age`. We use the *interval overlap* to measure the similarity of the logistic regression model confidence intervals between the confidential and synthetic data sets (Guo and Hu, 2022; Karr et al., 2006; Snoke et al., 2018):

$$IO = 0.5 \left\{ \frac{\min(u_c, u_s) - \max(l_c, l_s)}{u_c - l_c} + \frac{\min(u_c - u_s) - \max(l_c, l_s)}{u_s - l_s} \right\}, \tag{28}$$

where u and l denote upper and lower confidence interval bounds, and subscripts c and s denote the confidential and synthetic data. The maximum IO value is one, where higher IO values denote higher data utility.

In addition, we calculate the standardized difference (SD) in model coefficients (Snoke et al., 2018),

$$SD = |\hat{\beta}_c - \hat{\beta}_s| / \text{se}(\hat{\beta}_c) \tag{29}$$

where $\text{se}(\hat{\beta})_c$ is the estimated standard error of the coefficient from the confidential data. The minimum SD value is

zero, where lower SD values denote higher data utility.

Table 4 shows the median SD and IO across all coefficients and data sets for each synthesis method. The MNL based approach provides significantly more accurate coefficients and confidence intervals than every other method including the holdout data sets. This likely occurs since the coefficients of interest are also present in the MNL model used to synthesize the `state` variable. Applying the Inference prevention algorithm increases the SD and reduces the IO for both the MNL and CART based methods. The MNL - Inference data sets still provide similar or better utility than the holdout data set and any other synthesis method.

Synthesis Method	Median SD	Median IO
Holdout	1.32	0.63
MNL	0.79	0.81
MNL - Inference	1.21	0.70
CART	1.21	0.70
CART - Inference	1.96	0.45
MOSTLY.AI	2.13	0.54

TABLE 4 Standardized difference (SD) and confidence interval overlap (IO) results.

7.3 | Summary of Results

This application illustrates the ability of our proposed synthesis method to produce legally anonymous synthetic data, and reduce the probability of attribute disclosure for a sensitive binary variable while maintaining data utility. In terms of privacy, both the MNL, CART, and MOSTLY.AI synthesis approaches produced synthetic data that met the legal anonymity criteria. We found that attribute disclosure was still possible against data that met the legal anonymity criteria and that our Inference prevention algorithm successfully reduced the probability of inferring true sensitive values. While all proposed methods produced synthetic data that preserved spatial clustering, data utility as measured by the pMSE ratio and analysis specific utility was reduced by the algorithm. This tradeoff can likely be tuned by changing the value of c, although we did not experiment with this in our current application. Overall, the synthetic data produced by the MOSTLY.AI method could not match the utility of the synthetic data from the proposed methods. This could be due to lack of sufficient training data (Hradec et al., 2022), but this still represents an advantage for the proposed synthesis methods.

8 | DISCUSSION

Data synthesis is a promising and well established approach to protecting confidential data. It has been used to protect many types of data such as time series, structured data accompanying user-generated content, and location data. We showed that synthetic data sampled from the same data generating distribution as the confidential data is guaranteed to meet legal privacy criteria in expectation, and we proposed a new data synthesis approach that combined Bayesian optimization with commonly used data synthesis models to achieve this data synthesis. We found that the proposed synthesis approach is capable of generating legally anonymous data with reduced attribute disclosure risk and higher utility than a baseline comparison method which is known to produce legally anonymous data.

The proposed method is also flexible such that data owners could optimize the synthesis process using any single valued metric for utility or privacy rather than the pMSE ratio used in our empirical application. For example, the SPEC metric (Bowen et al., 2021) gives a single-valued measurement of distributional similarity based on the cumulative density functions of propensity scores. It would also be possible to optimize the synthesis process for analysis-specific utility. One example is to optimize the choice of κ in the synthesis method proposed by Schneider et al. (2018) to minimize the deviations in model parameters or maintain estimates of optimal markup percentages and profit ratios.

This paper does have some limitations. For example, we utilized a relatively small data set in our empirical application. Optimizing a synthesis model for a much larger data set would be more time intensive, but certainly feasible especially in the case of a one-off data release. We plan to study this in future work. Furthermore, the proposed Inference prevention algorithm is only applicable to a sensitive binary variable. Future work could consider generalizing the algorithm to handle attribute disclosure of multiple types of variables.

This paper demonstrates that legal anonymous data synthesis is possible using open source programming tools. We encourage further research into this promising area of disclosure limitation.

references

Anand, P. and Lee, C. (2023) Using deep learning to overcome privacy and scalability issues in customer data transfer. *Marketing Science*, **42**, 189–207.

Arthur, D. and Vassilvitskii, S. (2007) k-means++: The advantages of careful seeding. *Tech. rep.*

Asghar, H. J., Ding, M., Rakotoarivelo, T., Mrabet, S. and Kaafar, M. A. (2020) Differentially private release of high-dimensional datasets using the gaussian copula. *Journal of Privacy and Confidentiality*, **10**.

Azizi, Z., Zheng, C., Mosquera, L., Pilote, L. and El Emam, K. (2021) Can synthetic data be a proxy for real clinical trial data? a validation study. *BMJ open*, **11**, e043497.

Bale, C. D., Fischer, J., Schneider, M. J., Weber, S. and Chang, S. (2022) Legally anonymizing location data under the gdpr. *preprint*. URL: `https://www.researchgate.net/publication/361510709_Legally_Anonymizing_Location_Data_Under_the_GDPR`.

Bowen, C. M., Liu, F. and Su, B. (2021) Differentially private data release via statistical election to partition sequentially: Statistical election to partition sequentially. *Metron*, **79**, 1–31.

Bowen, C. M. and Snoke, J. (2021) Comparative study of differentially private synthetic data algorithms from the nist pscr differential privacy synthetic data challenge. *Journal of Privacy and Confidentiality*, **11**. URL: `https://journalprivacyconfidentiality.org/index.php/jpc/article/view/748`.

Brochu, E., Cora, V. M. and De Freitas, N. (2010) A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Deisenroth, M. P., Faisal, A. A. and Ong, C. S. (2020) *Mathematics for Machine Learning*. Cambridge University Press. URL: `https://mml-book.com`.

Drechsler, J. and Hu, J. (2021) Synthesizing geocodes to facilitate access to detailed geographical information in large-scale administrative data. *Journal of Survey Statistics and Methodology*, **9**, 523–548.

Drechsler, J. and Reiter, J. P. (2011) An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, **55**, 3232–3243.

El Emam, K., Mosquera, L., Jonker, E. and Sood, H. (2021) Evaluating the utility of synthetic covid-19 case data. *JAMIA open*, **4**, ooab012.

European Data Protection Board (2020) Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak. URL: https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-042020-use-location-data-and-contact-tracing_en.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2020) Generative adversarial networks. *Communications of the ACM*, **63**, 139–144.

Guo, S. and Hu, J. (2022) Data privacy protection and utility preservation through bayesian data synthesis: A case study on airbnb listings. *The American Statistician*, 1–9.

Hintze, M. and El Emam, K. (2020) Can synthetic data help organizations respond to 'schrems ii'? iapp. URL: https://iapp.org/news/a/can-synthetic-data-help-organizations-respond-to-schrems-ii/.

Hittmeir, M., Mayer, R. and Ekelhart, A. (2020) A baseline for attribute disclosure risk in synthetic data. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, 133–143.

Hradec, J., Craglia, M., Di Leo, M., De Nigris, S., Ostlaender, N. and Nicholson, N. (2022) Multipurpose synthetic population for policy applications. *EUR 31116 EN, Publications Office of the European Union, Luxembourg*.

Hu, J. (2019) Bayesian estimation of attribute and identification disclosure risks in synthetic data. *Transactions on Data Privacy*, **12**, 61–89.

Husnayain, A., Shim, E., Fuad, A. and Su, E. C.-Y. (2020) Assessing the community risk perception toward covid-19 outbreak in south korea: evidence from google and naver relative search volume. *medRxiv*.

Jackson, J., Mitra, R., Francis, B. and Dove, I. (2022) Using saturated count models for user-friendly synthesis of large confidential administrative databases. *Journal of the Royal Statistical Society: Series A*.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. and Sanil, A. P. (2006) A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, **60**, 224–232.

Kim, N. (2020) 'more scary than coronavirus': South korea's health alerts expose private lives. URL: https://www.theguardian.com/world/2020/mar/06/more-scary-than-coronavirus-south-koreas-health-alerts-expose-private-lives.

Koch, R. (2018) Data anonymization and gdpr compliance: the case of taxa 4x35. URL: https://gdpr.eu/data-anonymization-taxa-4x35,.

Lee, C. and Anand, P. (2020) Using deep learning to overcome privacy and scalability issues in customer data transfer. *Available at SSRN 3769521*.

Nogueira, F. (2014) Bayesian Optimization: Open source constrained global optimization tool for Python. URL: https://github.com/fmfn/BayesianOptimization.

Nowok, B., Raab, G. M. and Dibben, C. (2016) synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, **74**, 1–26.

Oganian, A. and Domingo-Ferrer, J. (2017) Local synthesis for disclosure limitation that satisfies probabilistic k-anonymity criterion. *Transactions on data privacy*, **10**, 61.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Platzer, M. and Reutterer, T. (2021) Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in big Data*, 43.

Quick, H. (2021) Generating poisson-distributed differentially private synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **184**, 1093–1108.

— (2022) Improving the utility of poisson-distributed, differentially private synthetic data via prior predictive truncation with an application to cdc wonder. *Journal of Survey Statistics and Methodology*, **10**, 596–617.

Quick, H., Holan, S. H., Wikle, C. K., and Reiter, J. P. (2015) Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spatial Statistics*, **14**, 439–451.

Quick, H., Holan, S. H. and Wikle, C. K. (2018) Generating partially synthetic geocoded public use data with decreased disclosure risk by using differential smoothing. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **181**, 649–661.

Reiter, J. P. (2005) Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, **21**, 441.

Reiter, J. P., Wang, Q. and Zhang, B. (2014) Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *Journal of Privacy and Confidentiality*, **6**.

Rubin, D. B. (1981) The bayesian bootstrap. *The annals of statistics*, 130–134.

Ruggles, S., Fitch, C., Magnuson, D. and Schroeder, J. (2019) Differential privacy and census data: Implications for social and economic research. In *AEA papers and proceedings*, vol. 109, 403–08.

Schneider, M. J. and Abowd, J. M. (2015) A new method for protecting interrelated time series with bayesian prior distributions and synthetic data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 963–975.

Schneider, M. J., Hu, J., Mankad, S. and Bale, C. D. (2023) Protecting the anonymity of online users through bayesian data synthesis. *Expert Systems with Applications*, **216**, 119409.

Schneider, M. J., Jagpal, S., Gupta, S., Li, S. and Yu, Y. (2018) A flexible method for protecting marketing data: An application to point-of-sale data. *Marketing Science*, **37**, 153–171.

Scikit-learn () sklearn.mixture.gaussianmixture. URL: `https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html`.

Snoek, J., Larochelle, H. and Adams, R. P. (2012) Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, **25**.

Snoke, J., Raab, G. M., Nowok, B., Dibben, C. and Slavkovic, A. (2018) General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society, Series A*, **181**, 663–688.

Taub, J., Elliot, M., Pampaka, M. and Smith, D. (2018) Differential correct attribution probability for synthetic data: an exploration. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings*, 122–137. Springer.

Wang, H. and Reiter, J. P. (2012) Multiple imputation for sharing precise geographies in public use data. *The annals of applied statistics*, **6**, 229.

Working Party (2014) Opinion 05/2014 on anonymisation techniques. URL: `https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf`.

Zastrow, M. (2020) South korea is reporting intimate details of covid-19 cases: has it helped? URL: `https://www.nature.com/articles/d41586-020-00740-y`.

9 | APPENDIX

9.1 | GMM and EM

Assuming the samples $\mathbf{x}_n,\ n \in [N]$ are i.i.d., the parameters $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^K$ that maximize the log-likelihood of observing $\mathbf{X}$ are obtained by solving the following problem:

$$\begin{aligned} \max_{\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^K} \quad & \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \text{s.t.} \quad & \sum_{k=1}^{K} \pi_k = 1 \\ & \pi_k \geq 0, \quad \forall k \end{aligned} \tag{30}$$

See Deisenroth et al. (2020) for a thorough explanation and derivation of the EM algorithm updates for GMMs. Let $r_{n,k}$ denote the probability that $\mathbf{x}_n$ was generated by the kth mixture component:

$$r_{n,k} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{31}$$

The mean, covariance, and mixture weight parameters are updated according to the following rules:

$$\boldsymbol{\mu}_k^{new} = \frac{1}{R_k} \sum_{n=1}^{N} r_{n,k} \mathbf{x}_n \tag{32}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{R_k} \sum_{n=1}^{N} r_{n,k} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T} \tag{33}$$

$$\pi_k^{new} = \frac{R_k}{N} \tag{34}$$

where $R_k = \sum_{n=1}^{N} r_{n,k}$. Given initial values for $\boldsymbol{\Theta}$, the EM algorithm alternates between calculating the responsibilities $r_{n,k}$ and updating $\boldsymbol{\Theta}$ until convergence. This will require computing the inverse of the covariance parameters, which may be singular if the mean of a component is identical to some $\mathbf{x}_n$ Deisenroth et al. (2020). In practice, this is handled by adding a small number to the diagonal of the covariance parameters Scikit-learn.