ORIGINAL ARTICLE

Legally Anonymous Sequential Synthesis of Location Data

Cameron D. Bale[1]

[1]Decision Sciences, Drexel University, Philadelphia, PA, 19103, USA

Correspondence
Cameron D. Bale, 3220 Market Street, Office 734, Drexel University, Philadelphia, PA, 19103, USA
Email: cdb327@drexel.edu

Funding information
None

Working abstract.

KEYWORDS
Working keywords.

1 | INTRODUCTION

Organizations that collect and disseminate sensitive data face a challenging task: they must maintain the utility of their data for analysis while protecting the identities and sensitive information of data subjects and ensuring compliance with privacy law. In general, to comply with privacy law, data controllers can either implement privacy and security measures for personal data (failing to do so can result in significant fines - for example, see Koch (2018)), or convert personal data to non-personal data, a process known as anonymization (European Data Protection Board, 2020). Anonymization is the conversion of sensitive or personal data into non-personal 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. The generation of synthetic data, which is analyzed or released in place of sensitive data, is one such method which is rapidly gaining popularity.

Under the General Data Protection Regulation (GDPR), anonymous data is described as "...information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable," (GDPR Recital 26 (European Parliament and Council of European Union, 2016)). A reasonability test is used to assess whether data is, in fact, anonymized under the GDPR. This reasonability test determines the robustness of anonymization using three criteria: (i) Singling out, (ii) Linkability, and (iii) Inference, which have been recently defined by Bale et al. (2022) for location trajectories anonymized using generalization and suppression.

Both academic researchers (e.g., (Quick, 2021)) and practitioners have realized the potential for synthetic data to provide a desirable balance between privacy and data utility. In 2021, there were at least twenty-two companies

providing methods for generating privacy preserving synthetic tabular data (Devaux, 2021). Recent predictions from Gartner state that "By 2025, synthetic data will reduce personal customer data collection, avoiding 70% of privacy violation sanctions," (Panetta, 2021), and that "By 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated," (White, 2021). To produce synthetic data, a data controller specifies a generative model which is intended 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).

Many examples of synthetic data generation have shown the potential to produce highly useful data with reduced disclosure risks. 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). The report claimed 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 passes certain privacy tests) (pg. 44). One of the use cases in the report involved generating hundreds of thousands of synthetic cancer patient records. The company MOSTLY.AI provided the commercial technology used for the data synthesis, complete with privacy metrics used to verify the anonymity of the synthetic records. This company directly states that their synthetic data is exempt from data privacy legislation[1]. Other companies and researchers also claim that synthetic data does not directly relate to, or identify real individuals, and is therefore not subject to privacy legislation. For example, Hintze and El Emam (2020) 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.

The goal of this paper is to study data synthesis as a legal anonymization mechanism. Based on the privacy metrics used by MOSTLY.AI in the JRC report (Hradec et al., 2022), we seek to determine whether popular parametric and non-parametric synthesis models can produce legally anonymous synthetic data. We show that the privacy metrics are guaranteed to be met in expectation when synthetic data is drawn from the same data generating distribution as the sensitive data. In our empirical application, we illustrate a new sequential data synthesis approach in which synthesis model parameters are optimized using the propensity score mean-squared error ratio to produce synthetic data with the required levels of privacy. We assess our sequential synthesis approach by applying it to a data set containing marked locations of COVID-19 patients in South Korea, and use the publicly available data synthesis tool from MOSTLY.AI as a baseline for generating legally anonymous synthetic data.

The paper proceeds as follows. In Section 2, we provide a review of the relevant literature. Section 3 describes a motivating example using a mapping application built on the location data of South Korean COVID-19 patients, and discusses how releasing sensitive data can lead to harm for data subjects. In Section 4, we study the privacy measures used by MOSTLY.AI in determining whether synthetic data is legally anonymous, and show that synthetic data which is drawn from the same data generating distribution as the sensitive data is guaranteed to meet the privacy metrics in expectation. We verify our findings using a simulation study. Section 5 contains the application of our sequential synthesis strategy to the South Korean COVID-19 location data related to Section 3, and the comparison to the baseline offering from MOSTLY.AI. We conclude and discuss opportunities for further research in Section 7.

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

2 | LITERATURE REVIEW

Synthetic data models have been used to generate protected data for a wide variety of use cases. In marketing, synthetic point-of-sale data can be used to protect store identities while enabling estimation of price elasticities and promotion effects (Schneider et al., 2018a). Synthetic data can enable secondary analysis of clinical trial data (Azizi et al., 2021) and research on sensitive patient datasets (El Emam et al., 2021). Synthetic data has also been used to protect time series (Schneider and Abowd, 2015) and fine-grained location data (Drechsler and Hu, 2021). There have even been attempts to infuse differential privacy guarantees into synthetic data (Bao et al., 2021; Quick, 2021).

Previous to this work, CARTs and Dirichlet Process mixtures of products of multinomials (DPMPMs) had been used by Drechsler and Hu (2021); Wang and Reiter (2012) to synthesize precise geographical data for the release of partially synthetic data. A method for generating fully synthetic location data with corresponding individual attributes has been proposed by Quick et al. (2015), but the proposed method is highly computationally intensive. And, while the previously cited works, in addition to Quick et al. (2018), examine the disclosure risks inherent in releasing synthetic data, none take steps to examine the specific risks described under any privacy laws, or bound these risks to a legally acceptable level. Thus, the methodological objective of this work is to examine under what conditions synthetic data is legally anonymous, and to propose a sequential synthesis strategy capable of producing legally anonymous synthetic data.

3 | MOTIVATING EXAMPLE: DATA SCIENCE FOR COVID-19 (DS4C)

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 help increase the public's trust and enable 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, can also help the public know where they may have been exposed to COVID-19.

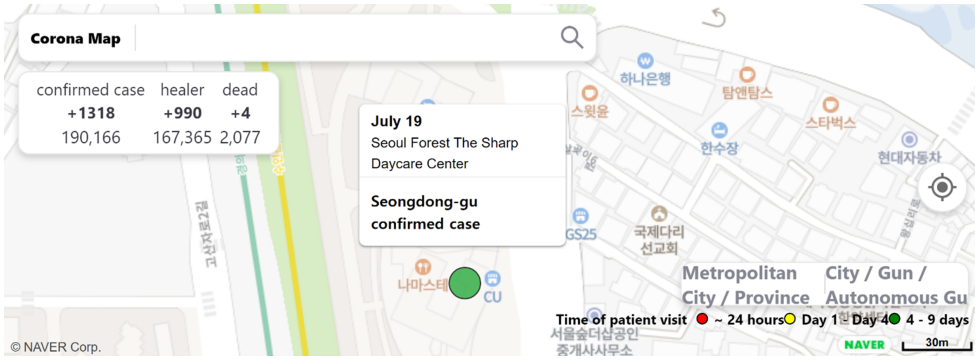

FIGURE 1 Screenshot from coronamap.site.

Unfortunately, even well-intentioned releases of sensitive data can harm data subjects. Feelings of privacy invasions may dissuade individuals from getting tested, and non-infected individuals may avoid places that an infected person was known to have visited long after it is safe to return (Zastrow, 2020). There can also be severe conse-

quences 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[2]. 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-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 examine whether a legally anonymous synthetic version of the data could be released which maintains much of the utility of the original data.

4 | PRIVACY CRITERIA

Consider two confidential data sets, $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_N)^\mathsf{T}$ and $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_N)^\mathsf{T}$, both of which contain N records drawn i.i.d. from the same data generating distribution $\mathcal{D}$. Suppose a data controller is interested in releasing a synthetic version of the data set $\mathbf{T}$, denoted $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)^\mathsf{T}$. We will refer to the data sets $\mathbf{T}$, $\mathbf{H}$, and $\mathbf{Z}$ as the training, holdout, and synthetic data sets, respectively. The intuition for the privacy criteria is as follows: synthetic data provides adequate privacy if, on average, it is not more similar to the training data than a second confidential data set from the same data generating distribution. In the words of Platzer and Reutterer (2021), "[this] allows to make a strong case for plausible deniability for any [record], as the synthetic data records do not allow to conjecture whether an individual was or was not contained in the training [data set]. Even for cases of a strong resemblance of a particular record with a real-world subject, it can be argued that such resemblance can occur for unseen subjects just as well," (pg. 6).

Per the synthetic data report given by MOSTLY.AI, there are three privacy criteria which must be met in a legally anonymous synthetic data set. First, the *identical match share*, or the proportion of records in the synthetic data set that are identical to records in the training data set must not be larger than the proportion of holdout records which are identical to training records. We let $d^z_{i,j} = \text{dist}(\mathbf{z}_i, \mathbf{t}_j)$ and $d^h_{i,j} = \text{dist}(\mathbf{h}_i, \mathbf{t}_j)$ denote distance measures between the ith and jth synthetic and training and holdout and training records, respectively. Further, let $d_i^{z(k)}$ and $d_i^{h(k)}$ denote the distance between the ith synthetic and holdout records and their kth nearest neighbor training record. The identical match share for the synthetic data set is defined as

$$IMS_{\mathbf{Z}} = \frac{1}{N} \sum_{i=1}^{N} [d_i^{z(1)} \leq \delta_{IMS}], \tag{1}$$

and the identical match share for the holdout data set is defined as

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

$$IMS_{\mathbf{H}} = \frac{1}{N}\sum_{i=1}^{N}[d_i^{h(1)} \leq \delta_{IMS}], \tag{2}$$

where $[.]$ are Iverson brackets. For records with continuous attributes, there is a zero probability of having an exactly identical record in a finite population. So, if using a distance measure applied to continuous attributes, 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 small bins, and a measure such as the Hamming distance may be used, in which case δ_{IMS} may be set to zero.

To be considered legally anonymous, the identical match share of the synthetic data set can be no more than that of the holdout data set,

$$IMS_{\mathbf{Z}} \leq IMS_{\mathbf{H}}, \tag{3}$$

which represents the first privacy criterion.

The next criterion is calculated from the distributions of nearest neighbor distances between the synthetic and training records, and the holdout and training records. Let $\mathbf{F}_{DCR}^{z(k)}$ and $\mathbf{F}_{DCR}^{h(k)}$ denote the empirical cumulative distribution functions of the distances between each synthetic record and its kth nearest neighbor, and each holdout record and its kth nearest neighbor, respectively. The *distance to closest record* criterion requires that the fifth percentile of the nearest neighbor distances from synthetic to training records cannot be smaller than the fifth percentile of the nearest neighbor distances from holdout to training records. The fifth percentiles of $\mathbf{F}_{DCR}^{z(k)}$ and $\mathbf{F}_{DCR}^{h(k)}$ satisfy the following,

$$\mathbf{F}_{DCR}^{z(1)}(p_{DCR}^{z}) \leq 0.05, \tag{4}$$

and

$$\mathbf{F}_{DCR}^{h(1)}(p_{DCR}^{h}) \leq 0.05, \tag{5}$$

where p_{DCR}^{z} and p_{DCR}^{h} are the fifth percentiles. For the synthetic data set to be considered legally anonymous, the fifth percentile of the distance to closest records for the synthetic data set can be no less than that of the holdout data set,

$$p_{DCR}^{z} \geq p_{DCR}^{h}, \tag{6}$$

which is the second criterion.

The final criterion is calculated using a ratio of nearest neighbor distances. The *nearest neighbor distance ratio*, denoted $NNDR_i^z$ for the ith synthetic record and $NNDR_i^h$ for the ith holdout record, is defined as

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

and

$$\frac{d_i^{h(1)}}{d_i^{h(5)}} = NNDR_i^h. \tag{8}$$

For a given synthetic or holdout record, this is the ratio of the distance to the nearest training record to the distance to the fifth nearest training record. Let $\mathbf{F}_{NNDR}^{z(k)}$ and $\mathbf{F}_{NNDR}^{h(k)}$ denote the empirical cumulative distribution functions of the nearest neighbor distance ratios of the nearest neighbor distance to the kth nearest neighbor distance for the synthetic and holdout data sets, respectively. The *nearest neighbor distance ratio* criterion requires that the fifth percentile of the nearest neighbor distance ratios for synthetic records cannot be smaller than the fifth percentile of the nearest neighbor distance ratios for holdout records. Letting p_{NNDR}^z and p_{NNDR}^h denote the corresponding percentiles, we have

$$p_{NNDR}^z \geq p_{NNDR}^h, \tag{9}$$

which is the third criterion.

Now that we have reviewed the legal privacy requirements, let us assume that the synthetic data set $\mathbf{Z}$ is actually drawn i.i.d. from a data synthesis model that captures the data generating distribution $\mathcal{D}$. This assumption is important because it guarantees that the above privacy requirements are met in expectation.

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 training 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 training records, each of which are drawn i.i.d. from $\mathcal{D}$. The random variable denoting the number of training 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 and p_C[3],

$$\begin{aligned} M_i &\sim Bin(N, p_C), \\ P(M_i = m_i) &= \binom{N}{m_i} p_C^{m_i} (1 - p_C)^{N - m_i} \end{aligned} \tag{10}$$

The expected number of identical records for a given synthetic record is then $N \times p_C$, which scales with the number of records. Intuitively, the larger the training data set, the more likely there is to be an identical match between a training record and a given synthetic record. For the purposes of identical match share, we are interested in the probability that there is at least one training record that is identical to each synthetic record, i.e., for a given synthetic record, we are interested in

[3] Binomial distribution assumes the N records are drawn with replacement from the population. Need to think about whether this assumption is valid.

$$\begin{aligned} P(M_i \geq 1) &= 1 - P(M_i = 0) \\ &= 1 - \prod_{j=1}^{N} P(C_{i,j} = 0) \\ &= 1 - (1 - P_C)^N \\ &= p_{IMS}. \end{aligned} \tag{11}$$

Define another Bernoilli random variable $\mathcal{M}_i = 1$ with probability p_{IMS} if the ith synthetic record has at least one match in the training data, and $\mathcal{M}_i = 0$ with probability $1 - p_{IMS}$. Then the random variable $\mathcal{M} = \mathcal{M}_1 + \mathcal{M}_2 + \cdots + \mathcal{M}_N$ is the total number of synthetic records which are identical to a training record and is Binomially distributed with expected value $N \times p_{IMS}$. The identical match share from (1) is calculated as

$$IMS_{\mathbf{Z}} = \frac{1}{N}\mathcal{M}. \tag{12}$$

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

$$\begin{aligned} E[IMS_{\mathbf{Z}}] &= \frac{1}{N}E[\mathcal{M}] \\ &= \frac{1}{N}N \times p_{IMS} \\ &= p_{IMS}. \end{aligned} \tag{13}$$

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 (3) is met in expectation for synthetic data sets drawn from the same data generating distribution as the training 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 $\mathcal{D}$.

4.1 | Estimating Distributional Similarity Using pMSE Ratios

Synthetic data generated using a synthesis model that correctly approximates the data generating distribution will meet the privacy criteria in Section 4 in expectation. Therefore, we require a method for determining distributional similarity between synthetic and training data sets.

Recent work from Snoke et al. (2018) demonstrated the use of the propensity score mean-squared error (pMSE) as a global utility measure for synthetic data. For a combined synthetic and confidential data set $\mathbf{D} = [\mathbf{Z}, \mathbf{T}]^{\mathsf{T}}$, the pMSE is estimated as

$$pMSE = \sum_{i=1}^{N} (\hat{p}_i - c)^2, \tag{14}$$

where N is the number of records in $\mathbf{D}$, c is the proportion of synthetic records in $\mathbf{D}$, and $\hat{\rho}_i$ is the probability that the ith record is synthetic which is estimated from some discriminatory model such as logistic regression or a CART model.

Consider the case where the synthetic data $\mathbf{Z}$ is drawn from a model which correctly approximates the distribution of the original data $\mathbf{T}$. Under this null case, there will still be some estimation error which arises from the propensity score model, and Snoke et al. (2018) analytically derive the expected value of the pMSE under this null case when logistic regression is used as the propensity score model. However, we desire synthetic data which is drawn from the same distribution as the training data, not the distribution of the training data itself. Snoke et al. (2018) provide the expected pMSE from a logistic regression for two synthetic data sets drawn from the same data synthesis model,

$$E[pMSE] = 2 * (k-1)(1-c)^2 c/N, \tag{15}$$

where k is the number of parameters in the logistic regression which correspond to synthetic variables. This is twice the expected pMSE under the previous null case.

Alternatively, a CART model can be used as the propensity score model, but the expected pMSE cannot be derived analytically. Instead, the null pMSE for synthetic data sets from the same data generating distribution can be estimated using the mean pMSE from a CART propensity score model applied to multiple pairs of synthetic data sets. The pMSE ratio for a given synthetic data set $\mathbf{Z}_m$ is the ratio of the pMSE between a synthetic and training data set to the null pMSE between two data sets from the same data generating distribution,

$$\mathcal{R}_m = pMSE_m / E[pMSE]. \tag{16}$$

When $\mathcal{R}_m$ is close to one, it indicates that the pMSE between the training and synthetic data is close to what is expected between data sets drawn from the same distribution, which is the desired case for meeting the privacy criteria in Section 4.

4.2 | Simulation Study

We perform a simulation study to verify that the privacy criteria in equations (3), (6), and (9) are met when $\mathbf{T}$, $\mathbf{H}$, and $\mathbf{Z}$ are sampled i.i.d. from the same data generating distribution $\mathcal{D}$. We also verify that the pMSE ratio based on the expected pMSE under the null of data generated from the same distribution is equal to one.

Our simulation study is similar to the one performed by Snoke et al. (2018). For a given simulation, we sample ten training, ten synthetic, and ten holdout data sets containing 5000 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 training/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, fifth percentile of the DCR, and fifth percentile of the NNDR for the ith training data set and the ith synthetic and holdout data sets for $i = 1, \ldots, 10$. We also calculate the pairwise pMSE ratios between the ith data sets using a logistic regression with main effects and interactions only, and a CART model using the mean pMSE between pairwise synthetic/holdout data sets to approximate the expected pMSE under the null. We perform 1000 simulations in total, and provide the means of the privacy measures and pMSE ratios under each different covariance value in Table

The simulation results in Table 1 confirm the analytical results from Section 4. When the synthetic data is drawn

Covariance	IMS Ratio	DCR Ratio	NNDR Ratio	LR pMSE Ratio	CART pMSE Ratio
0.0	1.008	0.998	0.999	0.976	1.000
0.1	1.000	1.000	1.001	0.974	0.997
0.2	1.009	0.999	0.999	1.012	1.003
0.3	0.998	1.001	0.999	0.995	1.001
0.4	0.996	1.001	1.001	1.020	0.999
0.5	0.997	1.001	1.000	1.004	1.003
0.6	1.000	0.999	0.999	0.994	1.004
0.7	1.002	0.998	0.999	1.004	1.001
0.8	1.000	0.999	0.999	0.993	0.983
0.9	1.000	1.001	1.001	0.993	0.996

TABLE 1 Results from simulation study. All privacy measures (IMS, DCR, and NNDR ratios) are calculated as the ratio of the holdout value to the synthetic value. The pMSE ratios are calculated based on the synthetic data relative to the training data.

from the same data generating distribution as the holdout data (and the training data) then the identical match share and DCR and NNDR percentiles are approximately equal in expectation between the synthetic and holdout data sets. We also confirmed the results from Snoke et al. (2018): the expected pMSE ratio between synthetic data and training data, when generated from the same distribution, is approximately one using the logistic regression and CART as propensity score models.

5 | ILLUSTRATIVE EXAMPLE: SOUTH KOREAN COVID-19 PATIENT LOCATION DATA

For our empirical application we consider a data set 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.

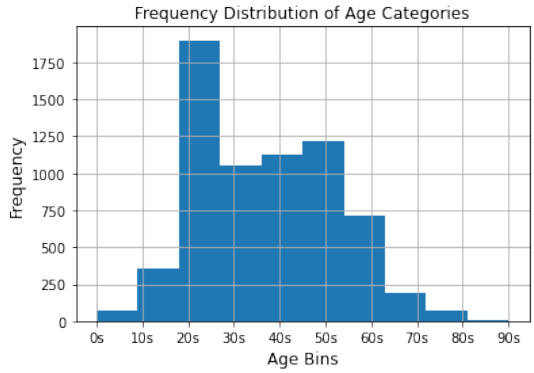

FIGURE 2 Frequency distribution of age variable.

5.1 | Data Synthesis Models

In this section, we present the models used for generating the synthetic data sets $\mathbf{Z}^{(l)}$. In the first synthesis step, we use a Gaussian mixture model (GMM) to approximate the bivariate location density[4]. This model is selected since it can accurately estimate the density of multi-modal distributions, which is important for sparsely distributed geographical locations.

Consider the dataset $\mathbf{Y}^{\ell} = [\mathbf{y}_1^{\ell}, \ldots, \mathbf{y}_N^{\ell}]^{\mathsf{T}}$, consisting of the location measurements of each record in $\mathbf{Y}$, denoted $\mathbf{y}_n^{\ell} = (y_{n1}, y_{n2})$. Using a GMM with K mixture components, the probability density function $p(\mathbf{y}^{\ell})$ is estimated as:

$$p(\mathbf{y}^{\ell}) = \sum_{k=1}^{K} \pi_k N(\mathbf{y}^{\ell}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{17}$$

Here, $N(\mathbf{y}^{\ell}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 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 $p(\mathbf{y}^{\ell})$ is a proper probability density. We assume our confidential locations $\mathbf{y}_n^{\ell}$, $n \in [N]$ to be i.i.d., and obtain the parameters $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^{K}$ that maximize the log-likelihood of observing $\mathbf{Y}^{\ell}$ by solving the following problem using expectation maximization (EM)[5]:

$$\begin{aligned} \max_{\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^{K}} \quad & \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k N(\mathbf{y}_n^{\ell}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \text{s.t.} \quad & \sum_{k=1}^{K} \pi_k = 1 \\ & \pi_k \geq 0, \quad \forall k. \end{aligned} \tag{18}$$

Once we obtain $\boldsymbol{\Theta}^*$, we sample N synthetic locations to create the synthetic location dataset $\mathbf{Z}^{\ell} = (\mathbf{z}_1^{\ell}, \ldots, \mathbf{z}_N^{\ell})$.

We compare two sequential synthesis strategies for the remaining categorical variables. The full synthesis process is as follows. First, we sample N synthetic locations from the GMM obtained by solving the maximization problem

[4] See the appendix for a review on the use of GMMs in the data privacy literature.

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

in (18). The first sequential synthesis strategy uses a categorical CART to approximate the marginal probability mass function for *sex* conditional on the confidential locations $\mathbf{Y}^{\ell}$. We then sample N synthetic *sex* values from the CART model conditional on the synthetic locations $\mathbf{Z}^{\ell}$. This process is repeated using categorical CART models for *age* and *state* as well. The second sequential synthesis strategy uses L1-penalized multinomial logit models with up to three-way powers and interactions for the sequential synthesis of *sex*, *age*, and *state*.

5.2 | Bayesian Optimization of Synthesis Models

When generating synthetic data, a data controller will typically want to maximize the utility of the synthetic data while meeting some minimum requirements for privacy. Synthesis models are often designed with parameters to directly control the privacy-utility trade-off. For example, see (Schneider et al., 2018b). The downside to the sequential synthesis strategy is that multiple synthesis models must be tuned to give a desirable privacy-utility trade-off, increasing the difficulty of implementation on the part of a data controller.

We propose the use of Bayesian optimization based on a Gaussian process as a solution to this problem (Snoek et al., 2012). This technique is a popular choice for choosing the optimal hyper-parameters for black-box machine learning models. In general, this technique is used to optimize an objective function without a closed form expression, but from which we can observe the function at sampled values. A thorough review of the Bayesian optimization technique is given by Brochu et al. (2010).

We use the pMSE ratio described in Section 4.1 as the main component of our objective function. To estimate the pMSE ratio, we use a CART model to discriminate between the confidential and synthetic records in a combined data set $\mathbf{D} = [\mathbf{Z}, \mathbf{T}^{(l)}]^{\mathsf{T}}$. For expected pMSE, we use the mean pairwise pMSE between each of twenty synthetic data sets.

To ensure that our sequential synthesis strategy produces synthetic data sets that meet the privacy criteria in Section 4 with good utility, we use the Bayesian optimization algorithm to select the synthesis models' hyperparameters to minimize the squared deviation of the average pMSE ratio from 1, across m synthetic datasets,

$$\min_{\{\boldsymbol{\theta}\}} \quad (1 - \frac{1}{m}\sum_{i=1}^{m} \text{pMSE}_m / E[\text{pMSE}])^2. \tag{19}$$

We use $\boldsymbol{\theta} = \{K, \alpha_{age}, \alpha_{sex}, \alpha_{state}\}$ to denote the number of components K for the GMM and the cost-complexity tuning parameter α for each of the categorical CARTs, and $\boldsymbol{f}$ to denote the objective function in (19)[6]. When synthesizing using the multinomial logit models, we optimize a parameter C for each model which denotes the inverse of the regularization strength.

We define $\boldsymbol{\theta}_i$ as the ith sample of synthesis model hyper-parameters, where the objective $\boldsymbol{f}(\boldsymbol{\theta}_i)$ gives the measure of data utility at $\boldsymbol{\theta}_i$. 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 $\boldsymbol{f}(\boldsymbol{\theta}_i)$.

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

[6] Note that the Gaussian process treats K as a continuous random variable, but we convert K to an integer when training the GMM to evaluate the pMSE.

random points, and then performs 25 iterations. Based on the θ in each iteration, we evaluate 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 Figures 3 and 4. We select the θ^* with the best overall f value, shown for each synthesis strategy in Table 2.

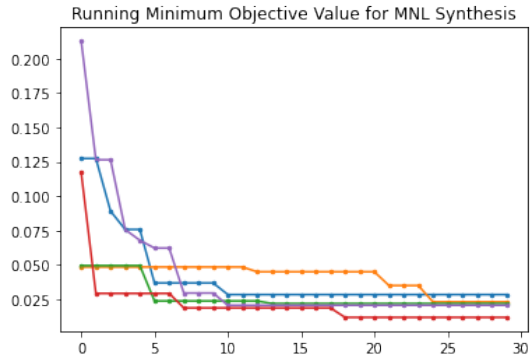

FIGURE 3 Running minimum of objective f for the multinomial based synthesis.

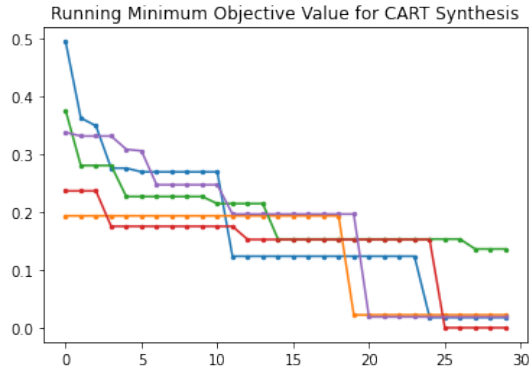

FIGURE 4 Running minimum of objective f for the CART based synthesis.

Based on the optimization process, both synthesis strategies were able to obtain objective values relatively close to zero, with the CART based strategy producing an average pMSE ratio closer to one. The strategies selected similar numbers of mixture components. The rankings of the regulariztion and complexities selected for each categorical variable are consistent across synthesis strategies as well. For *age*, the MNL synthesizer applied the least regularization, and the CART synthesizer has the smallest complexity parameter. Stronger regularization (less complexity) is used when synthesizing *sex* relative to *age*, and the strongest regularization (least complexity) is used when synthesizing *state*.

Variable	GMM + MNL	GMM + CART
Objective	$f = 0.0120$	$f = 0.0001$
Lat/Long	$K = 245$	$K = 293$
Sex	$C = 0.311$	$\alpha = 0.003$
Age	$C = 0.452$	$\alpha = 0.001$
State	$C = 0.189$	$\alpha = 0.054$

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).

6 | 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 training and holdout and training data sets with $\delta = 0.001$. Prior to calculating the distances, we standardize the training data set, and standardize the synthetic and holdout data sets using the means and standard deviations from the training data.

6.1 | Privacy Criteria

Synthesis Type	IMS (Synthetic - Holdout)	DCR (Synthetic - Holdout)	NNDR (Synthetic - Holdout)
MNL Based	-0.4789	0.0010	0.0238
CART Based	-0.4611	0.0008	0.0222
MOSTLY.AI	-0.3993	0.0000	0.0247

TABLE 3 Privacy results.

Table 3 contains the average differences in the privacy measures between the twenty synthetic data sets for each synthesis method and the holdout data set. The negative values for the IMS difference indicate that there are significantly more holdout than synthetic records which are identical to training records. Of the three synthesis methods, the MOSTLY.AI synthesis strategy produces the highest proportion of identical synthetic records, on average. The proportion of identical synthetic records produced by any of the synthesis methods is significantly less than what is contained in the holdout data set, thereby meeting the privacy requirement. The differences in the fifth percentile of the DCR distribution are small, with the MOSTLY.AI strategy producing the smallest difference once again. However, none of the DCR distributions for the synthetic data (as measured by the fifth percentile) are closer to zero than the holdout data, indicating that the synthesis methods do not produce synthetic records that are systematically closer to the training data than the holdout data is to the training data. Similarly, the fifth percentile of the NNDR distribution for the holdout data is smaller than for the NNDR distributions for the synthetic data, i.e., holdout points tend to be closer to their nearest neighbors relative to their fifth nearest neighbors compared to the synthetic data from all

methods. In general, the privacy results for IMS and DCR indicate the strongest privacy for the MNL based strategy, followed by the CART-based and MOSTLY.AI methods. On the other hand, the smallest differences in the fifth percentile of the NNDR ratio occur under the CART synthesis strategy, and the largest occur under the MOSTLY.AI synthesis strategy. This indicates a slightly higher tendency for the CART and MNL based strategies to produce synthetic points closer to their nearest training neighbors relative to their fifth nearest training neighbors, relative to the MOSTLY.AI strategy. Overall, all synthesis strategies meet the privacy criteria on average across the twenty synthetic data sets.

6.2 | Utility Measures

Next, we assess the utility of the synthetic data sets using three measures of data utility: (1) The pMSE ratio described in Section 5.2, which was used in the objective of the bayesian optimization process; (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; and (3) the average total variational distance between one-, two-, and three-way cross tabulations of all variables (Platzer and Reutterer, 2021).

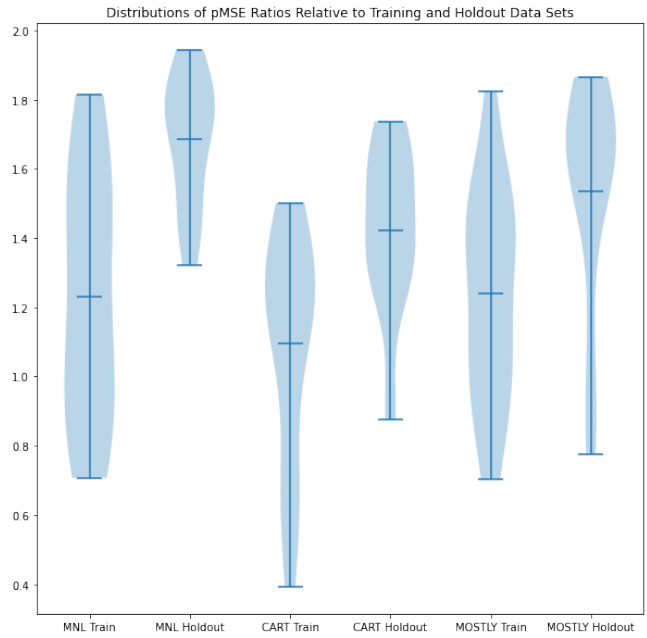

FIGURE 5 Distributions of pMSE ratios calculated using a CART model with null given by mean of pMSE under pairwise synthetic data sets.

For a baseline, we calculate the pMSE ratio between the training and holdout data sets using a logistic regression with main effects and interactions, and use the expected value of the pMSE under the null of data sets generated from the same synthesis model. This ratio comes out to 1.04, indicating a high distributional similarity between the holdout and training data sets.

Figure 5 shows the distributions of the pMSE ratios estimated using CART models for the synthetic data of each

synthesis method relative to the training data set and the holdout data set. In general, the pMSE ratios of the synthetic vs. holdout data tend to be higher than those of the synthetic vs. the training data, indicating more distributional similarity to the training data. The difference is most pronounced under the MNL strategy. The synthetic data sets with the highest utility are produced under the CART based strategy, as the average pMSE ratio under this method is 1.09. The utility of the MNL-based and MOSTLY.AI methods is slightly lower, producing average pMSE ratios of 1.23 and 1.24, respectively. On average, none of the synthesis methods produce synthetic data sets which are more similar to the training data than the holdout data is to the training data as measured by the propensity score models.

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. An estimate of this function can be obtained using the following:

$$\hat{K}(h) = \frac{|D|}{N} \sum_{n=1}^{N} \sum_{m=1}^{N} I(||\mathbf{y}_n^{\ell} - \mathbf{y}_m^{\ell}|| \leq h)/N, \text{ for } m \neq n. \tag{20}$$

Here, $I(.)$ is the indicator function which is equal to one whenever points $\mathbf{y}_n^{\ell}$ and $\mathbf{y}_m^{\ell}$ 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:

$$\hat{L}(h) = \sqrt{\hat{K}(h)/\pi} - h, \tag{21}$$

where positive values of this estimate indicate spatial clustering. We estimate the values of the L function for the training data and each of the $m = 20$ synthetic data sets 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 datasets. These results are shown in Figure 6. The locations synthesized from the GMM are extremely similar to the original locations based on the L-function values. Under the MNL and CART synthesizers, the locations with a *state* value of one (passed away from the disease), exhibit less spatial clustering than the original data. On the other hand, the MOSTLY.AI synthesizer produces an overall location distribution with less spatial clustering than the original data, but better preserves the spatial clustering for points with *state* equal one than the other synthesizers.

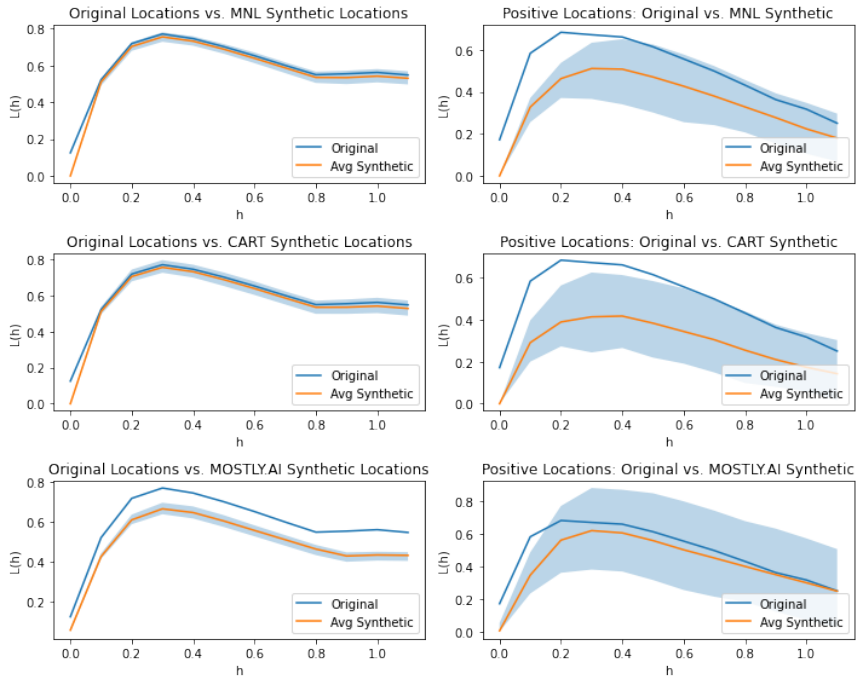

FIGURE 6 L-functions for the confidential data vs. average L-function values (orange) and 95% empirical confidence intervals (shaded) across synthetic data sets for each synthesis method.

The L-function results are confirmed when we examine the distributions of the training and holdout locations relative to the synthesized locations. We selected one set of synthetic locations from the GMM and one set from the MOSTLY.AI method for comparison to the training and holdout data sets. The locations synthesized from the GMM bear a strong visual resemblance to the training and holdout data, while the synthetic locations from the MOSTLY.AI synthesizer are more dispersed.

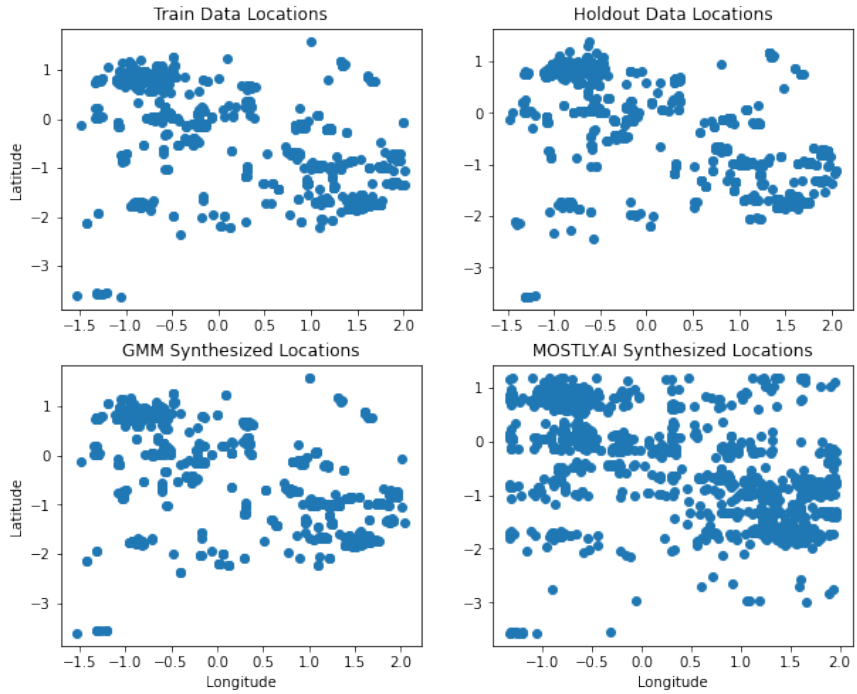

FIGURE 7 Comparison of training, holdout, and synthetic locations.

Our third and final measure of utility consists of measuring the average total variational distance between one-, two-, and three-way cross tabulations of all variables in the synthetic data sets compared to the training data set (Platzer and Reutterer, 2021). For a given data set with m variables, there are $\binom{m}{k}$ different k-way interactions. The total variational distance for a given interaction i is given by

$$TVD_i(\mathbf{T},\mathbf{Z}) = \frac{1}{2}\sum_x |f_i^T(X=x) - f_i^Z(X=x)|, \tag{22}$$

where f_i^T and f_i^Z denote the empirical distributions of the interaction i within the training and synthetic data sets, respectively, and x is one of the interaction values. Then the utility of the synthetic data set $\mathbf{Z}$ relative to the training data set $\mathbf{T}$ is given by

$$U^k(\mathbf{T},\mathbf{Z}) = 1/\binom{m}{k} \times \sum_i TVD_i(\mathbf{T},\mathbf{Z}), \tag{23}$$

which is the average total variational distance across all k-way interactions. This measure is highly similar to the deviations in the relative frequencies of cross-tabulations for the synthetic and confidential data (Drechsler and Hu, 2021).

In Table 4 we calculate (23) for all one-, two-, and three-way interactions for each of the twenty data sets for

each synthesis method, and compute the ratio of these values to the average total variational distance between the holdout and training data sets. The closer the ratio is to one, the closer the variation in the marginal distributions to what is expected between two data sets from the same data generating distribution. A ratio less than one is indicative of a synthesis model that is over-fit to the training data. The MOSTLY.AI method produces one way marginals with the highest utility, but the CART-based method is best at approximating the two- and three-way marginals. Interestingly, the utility (relative to the holdout data) improves for the MNL- and CART-based methods as k increases, but the opposite occurs for the MOSTLY.AI method. Overall, even the synthetic data sets with the highest utility for any k-way interaction still have approximately 60-70% more deviation in the marginal distributions than the holdout data set.

Synthesis Method	1-way Interactions	2-way Interactions	3-way Interactions
MNL	1.753	1.722	1.715
CART	1.718	1.623	1.594
MOSTLY.AI	1.546	1.657	1.722

TABLE 4 Average ratios of average total variational distance for each interaction in the synthetic data sets relative to the holdout data sets.

7 | DISCUSSION

In this paper, we have shown that synthetic data drawn from the same data generating distibution as sensitive data is guaranteed to meet legal privacy requirements in expectation. We used these findings to inform a new Bayesian optimization based sequential synthesis approach which automates the privacy-parameter selection procedure for data controllers.

In our empirical application, we showed that synthetic locations of South Korean COVID-19 patients can be synthesized with privacy meeting the current legal baseline under the GDPR (Hradec et al., 2022). For this particular data set, the proposed synthesis approach using CART models met the legal privacy requirement with better utility than the baseline comparison. The utility and privacy analyses show that the synthetic data is no more similar to the training data than a holdout data set from the same data generating distribution, indicating a sort of 'plausible deniability' if any similarity exists between synthetic records and confidential records.

Some of the difference in performance between the sequential synthesis approach and the deep-learning based MOSTLY.AI approach may be due to the small data set that was used. Future work should compare the proposed approach to deep-learning based synthesis on a larger data set. It would also be of interest to assess the resilience of legally anonymized data against inferential privacy attacks, such as the correct attribution probability measure (Taub et al., 2018). Ultimately, legal privacy standards will need to adapt to changing privacy risks and synthesis methods to ensure that useful data can be shared and processed with reduced privacy risk to data subjects.

references

(2021) An important census product may soon use synthetic data. The Economist. URL: `https://www.economist.com/united-states/2021/06/24/an-important-census-product-may-soon-use-synthetic-data`.

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.

Bao, E., Xiao, X., Zhao, J., Zhang, D. and Ding, B. (2021) Synthetic data generation with differential privacy via bayesian networks. *Journal of Privacy and Confidentiality*, **11**.

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.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **Vol. 39**, 1–38. URL: https://www.jstor.org/stable/2984875.

Devaux, E. (2021) List of synthetic data startups and companies - 2021. URL: https://elise-deux.medium.com/the-list-of-synthetic-data-companies-2021-5aa246265b42.

Diao, X., Yang, W., Wang, S., Huang, L. and Xu, Y. (2020) Privgmm: Probability density estimation with local differential privacy. In *International Conference on Database Systems for Advanced Applications*, 105–121. Springer.

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.

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.

European Parliament and Council of European Union (2016) Regulation (eu) 2016/679 of the european parliament and of the council. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Ge, F., Ju, Y., Qi, Z. and Lin, Y. (2018) Parameter estimation of a gaussian mixture model for wind power forecast error by riemannian l-bfgs optimization. *IEEE Access*, **6**, 38892–38899.

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/.

Hosseini, R. and Sra, S. (2015) Matrix manifold optimization for gaussian mixtures. *Advances in Neural Information Processing Systems*, **28**, 910–918. URL: https://proceedings.neurips.cc/paper/2015/file/dbe272bab69f8e13f14b405e038deb64-Paper.pdf.

— (2020) An alternative to em for gaussian mixture models: batch and stochastic riemannian optimization. *Mathematical Programming*, **181**, 187–223.

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*.

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*.

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*.

Majumdar, A., Hall, G. and Ahmadi, A. A. (2019) A survey of recent scalability improvements for semidefinite programming with applications in machine learning, control, and robotics. *arXiv 1908.05209*.

McLachlan, G. J. and Rathnayake, S. (2014) On the number of components in a gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **4**, 341–355.

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

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

Panetta, K. (2021) You'll be breaking up with bad customers - and 9 other predictions for 2022 and beyond. URL: https://www.gartner.com/en/articles/you-ll-be-breaking-up-with-bad-customers-and-9-other-predictions-for-2022-and-beyond-.

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.

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.

Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, **Vol. 26**, 195–239. URL: https://www.jstor.org/stable/2030064.

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

— (2018b) 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.

Sra, S. and Hosseini, R. (2013) Geometric optimisation on positive definite matrices for elliptically contoured distributions. *Advances in Neural Information Processing Systems*, **26**, 2562–2570.

Taub, J., Elliot, M., Pampaka, M. and Smith, D. (2018) Differential correct attribution probability for synthetic data: an exploration. In *International Conference on Privacy in Statistical Databases*, 122–137. Springer.

Vanderbei, R. J. and Benson, H. Y. (2000) On formulating semidefinite programming problems as smooth convex nonlinear optimization problems. *Technical Report*.

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

Wang, S., Zhu, B., Li, C., Wu, M., Zhang, J., Chu, W. and Qi, Y. (2020) Riemannian proximal policy optimization. *arXiv 2005.09195*.

White, A. (2021) By 2024, 60% of the data used for the development of ai and analytics projects will be synthetically generated. URL: `https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/`.

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`.

Wu, Y., Wu, Y., Peng, H., Zeng, J., Chen, H. and Li, C. (2016) Differentially private density estimation via gaussian mixtures model. *IEEE/ACM 24th International Symposium on Quality of Service (IWQoS)*, 1–6.

Xu, L. and Jordan, M. I. (1996) On convergence properties of the em algorithm for gaussian mixtures. *Neural Comput.*, **8**, 129–151.

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`.

8 | APPENDIX

8.1 | Gaussian Mixture Models and Privacy

Gaussian mixture models are used for density estimation and classification tasks. One strength of GMMs is that they can accurately estimate the density of multi-modal distributions. This is our main motivation for using a GMM, since Releasing the values of GMM parameters may compromise the privacy of the data the model was fit to. Central and local differentially private GMMs have been created to alleviate this concern. In general, privacy protection based on central differential privacy involves a data controller, who possesses personal data, using a differentially private mechanism to release private statistics about the personal data. In the case of GMMs, a data controller estimates the joint density of personal data using a differentially private GMM and releases this model to the public, who can use the model to learn about the underlying data Wu et al. (2016). On the other hand, local differential privacy involves implementing protection (noise addition) at the point of data collection so that data controllers are never in possession of sensitive data. In this case, the authors in Diao et al. (2020) propose a variant of a GMM designed to estimate the joint

density of the original data based on the noisy protected data. Note that any calculation performed on the output of a differentially private mechanism is also differentially private. This is one of the main strengths of differential privacy. However, differential privacy is known to severely reduce data utility, especially for detailed microdata Ruggles et al. (2019), which motivates the use of synthetic data methods for anonymizing data. For example, the United States Census Bureau has announced that it will move toward using a statistical model, rather than differential privacy based methods, to release synthetic microdata from the American Community Survey Cen (2021).

GMMs have rarely been used for data synthesis, the exception being (Oganian and Domingo-Ferrer, 2017), who use a GMM to generate synthetic data to satisfy probabilistic *k*-anonymity. An alternative option is to use a Gaussian copula model. A differentially private Gaussian copula model has been used to generate synthetic mixed variable datasets that provide highly accurate answers to one-way and two-way positive conjunction queries Asghar et al. (2020). 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.

8.2 | Methods for Obtaining GMM Parameters

The largest barrier to obtaining the parameters of a GMM is the constraint that covariance matrices must be positive definite. Optimization methods such as Newton, quasi-Newton, and scoring methods can be used to obtain mixture parameters, but come with significant downsides in terms of computation cost and complexity Redner and Walker (1984).

The parameters of GMMs are typically found using the expectation maximization (EM) algorithm Dempster et al. (1977). EM has significant benefits such as automatically satisfying probabilistic constraints, exhibiting monotonic convergence without the need to set a learning rate, and requiring low computational overhead Xu and Jordan (1996). While the EM algorithm exhibits slow convergence toward optimal parameter values in problems where mixture components are poorly separated, it still produces rapid increases in likelihood Xu and Jordan (1996),Redner and Walker (1984). In problems where mixture components are poorly separated, other methods are likely to perform poorly as well Xu and Jordan (1996).

The semi-definite programming literature describes techniques for solving problems with a nature similar to a GMM - maximizing an objective with a positive definite matrix constraint. One general strategy for handling this constraint is to factor the positive definite matrix. Such factorizations leverage problem-specific properties such as sparsity or the ability to use a low-rank approximation of the matrix of interest Majumdar et al. (2019). These properties are not afforded to GMMs, since covariance matrices cannot be singular and may not be sparse. Still, some semi-definite programming techniques have been tested in solving for mixture model parameters and on related statistical problems.

One factorization proposed in Vanderbei and Benson (2000) enables semi-definite programming problems to be expressed as standard convex optimization problems. This strategy was found to be inferior to an 'EM-like' fixed point algorithm in terms of speed when solving a problem similar to a mixture model. This was due to the computational expense of each iteration Sra and Hosseini (2013).

The authors in Hosseini and Sra (2015) tested a conjugate gradient method in solving for the parameters of a GMM in its original form and with a Cholesky decomposition of the covariance matrices. The EM algorithm was significantly faster in most tests. However, these authors also proposed a reformulated version of the GMM that lends itself to Riemannian optimization. This reformulation, paired with manifold stochastic gradient descent, was shown to outperform the EM algorithm by achieving better log-likelihood values and faster convergence to an optimum in some cases.

The reformulated GMM solved via Riemannian optimization has already been applied in forecasting wind power Ge et al. (2018) and Markov decision process problems Wang et al. (2020). Due to this recent success, we chose to test both EM and Riemannian optimization in solving for the parameters of our GMM.

8.3 | 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:

$$\begin{aligned} \max_{\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^K} \quad & \sum_{n=1}^N \log \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ \text{s.t.} \quad & \sum_{k=1}^K \pi_k = 1 \\ & \pi_k \geq 0, \quad \forall k \end{aligned} \tag{24}$$

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:

$$r_{n,k} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{25}$$

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

$$\boldsymbol{\mu}_k^{new} = \frac{1}{R_k} \sum_{n=1}^N r_{n,k} \mathbf{x}_n \tag{26}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{R_k} \sum_{n=1}^N r_{n,k} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T} \tag{27}$$

$$\pi_k^{new} = \frac{R_k}{N} \tag{28}$$

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.

8.4 | Reformulated GMM and Riemannian Optimization

The interested reader should consult Hosseini and Sra (2020) for detailed explanations and derivations of the GMM reformulated for Riemannian optimization. A brief overview and essential equations are presented here. The log-likelihood presented in (18) must be reformulated to be geodesically convex to enable Riemannian optimization meth-

ods to compete with EM in solving for GMM parameters. First, augment all samples $\mathbf{x}_n$ with an extra dimension:

$$\mathbf{y}_n^\mathsf{T} = [\mathbf{x}_n^\mathsf{T} \quad 1] \tag{29}$$

Next, consider the following maximization problem, which consists of a single Gaussian, using the augmented samples $\mathbf{y}_n$:

$$\max_{\mathbf{S}} \quad \sum_{n=1}^{N} \log q_{\mathcal{N}}(\mathbf{y}_n; \mathbf{S}) \tag{30}$$

where $q_{\mathcal{N}}(\mathbf{y}_n; \mathbf{S}) = 2\pi \exp(\frac{1}{2})\, \mathcal{N}(\mathbf{y}_n; 0, \mathbf{S})$. Let $\mathbf{S}^*$ denote the maximum of (30). Solving for $\mathbf{S}^*$ yields the optimal solution to (18) when $K = 1$:

$$\mathbf{S}^* = \begin{bmatrix} \boldsymbol{\Sigma}^* + \boldsymbol{\mu}^* \boldsymbol{\mu}^{*\mathsf{T}} & \boldsymbol{\mu}^* \\ \boldsymbol{\mu}^{*\mathsf{T}} & 1 \end{bmatrix} \tag{31}$$

Overall, the original GMM problem is replaced with a reparameterized log-likelihood consisting of geodesically convex components with local maxima that agree with those of (18). Let $\omega_k = \log \frac{\pi_k}{\pi_K}$, where $\omega_K = 0$. Using this change of variables to incorporate the contraints in (18) produces the following optimization problem on the product manifold $(\Pi_{k=1}^{K} \mathbb{P}^{D+1}) \times \mathbb{R}^{K-1}$, where $\mathbb{P}^{D+1}$ is the manifold of $(D+1) \times (D+1)$ positive definite matrices:

$$\max_{\{\mathbf{S}_k\}_{k=1}^{K}, \{\omega_k\}_{k=1}^{K-1}} \quad \sum_{n=1}^{N} \log \sum_{k=1}^{K} \frac{\exp(\omega_k)}{\sum_{j=1}^{K} \exp(\omega_j)} q_{\mathcal{N}}(\mathbf{y}_n; \mathbf{S}_k) \tag{32}$$

In this paper, both problems are solved - (18) and (32) - for various values of K. The values of K that minimize the Bayesian Information Criterion of the models are selected McLachlan and Rathnayake (2014). Based on these values of K, we select the solver that produces a model that generates the synthetic data with spatial properties most similar to those of the original data.