Journal of Privacy and Confidentiality (2013) **5**, Number 1, 73–105

Differential Privacy Applications to Bayesian and Linear Mixed Model Estimation

John M. Abowd*, Matthew J. Schneider†, and Lars Vilhuber‡

1 Introduction

In this paper, we investigate two approaches for applying differential privacy to the estimation of Linear Mixed Models (LMMs) and Bayesian Linear Mixed Models (BLMMs). We contribute to the statistics literature by creating new methodologies that apply existing differential privacy approaches to mixed-effect models. Mixed-effect models are widely used when organizations need to estimate thousands of small groups or areas in one formal probability model. Mixed-effect models take advantage of sparse computational procedures and condition on a small number of variance parameters that are used in the estimation of the realized effects for small groups, which may or may not be hierarchical. No known differentially private algorithms exist for this class of models and we propose two approaches based on a LMM and a BLMM. The first approach constructs an efficient, differentially private estimator that converges in distribution to the Maximum Likelihood Estimator (MLE) by using a sub-sample and aggregate algorithm [25]. The second approach produces differentially private linear predictors for regularized Empirical Risk Minimization (ERM) by perturbing an objective function [6]. Our methods harmonize the two approaches using continuous data and make appropriate methodological decisions where theory is missing. For example, the differentially private linear predictors for ERM are classifiers and usually have a binary dependent variable, but we extend the approach to a continuous but bounded dependent variable.

The main contribution of this paper is not the design of an end-to-end differentially private workflow for data analysis in linear mixed models. Instead, our contribution is the evaluation of how much accuracy one could reasonably expect from differentially private techniques, such as sample-and-aggregate and objective-perturbation. Our actual implementations are *not* strictly differentially private because of practical considerations designed to improve the usefulness (utility) of the released statistics. For example, we use empirical bounds on the dependent variables, rather than strict theoretical bounds. We also insert a bias-reduction step in the sample-and-aggregate method. Finally, rather than search over all possible extreme values when implementing objective-perturbation, we examine outlier and influential points selected using conventional statistical criteria. None of these refinements is strictly differentially private because of their dependencies on the actual sample data. One interpretation of our results is that they are relatively

*Department of Economics and Labor Dynamics Institute, Cornell University, Ithaca, USA, mailto:john.abowd@cornell.edu.

†Department of Statistical Science and Labor Dynamics Institute, Cornell University, Ithaca, USA, mailto:mjs533@cornell.edu.

‡Department of Economics and Labor Dynamics Institute, Cornell University, Ithaca, USA, mailto:lars.vilhuber@cornell.edu.

optimistic. Another interpretation is that they are indicative of what could be achieved with additional effort in fine-tuning differentially private algorithms.

A randomized function K gives ϵ-differential privacy [11] if for all data sets D_1 and D_2 differing on at most one element and all measurable subsets $S \subseteq Range(K)$,

$$\Pr[K(D_1) \in S] \leq \exp(\epsilon) \times \Pr[K(D_2) \in S].$$

In our implementation, D_1 and D_2 differ by the deletion of a single row of D_1 to form D_2. Statistical disclosure limitation (SDL) and privacy-preserving data mining (PPD) share the common goal of permitting an analyst to draw valid inferences about the properties of confidential data without revealing too much to the analyst about specific entities in the database. One approach to detailed tabulations is to collect data on a sufficiently large number of entities so as to ensure that no published number is based on only a few. As informal as this approach sounds, it lies at the heart of most of the disclosure limitation protocols in use by government agencies around the world, and its formalization as identity risk control (in SDL) and k-anonymity (in PPD) provides the basis for many rule-driven publication programs [9]. An alternative approach to protecting the confidentiality of the data provided by the entities that inhabit the levels of a detailed factor is the model-based approach. Model-based procedures combine the data from all of the entities using a formal probability model. The estimate for a particular level of the detailed factor is a function of all of the data [9], [11].

The linear mixed-effect model is a canonical form of interest to many because it is the basis for applied work in a wide variety of physical and social sciences. In addition, and perhaps of more interest in our context, it is the statistical workhorse of small-area estimation, which is an important part of many statistical agencies' publication programs [16], [21], [22]. Small-area estimation and its counterpart in economic data—detailed industrial tabulation—attempt to estimate regression-adjusted means for classifications that have many levels and are sparsely represented in the underlying confidential data. In linear mixed-effect models, the analyst is often interested in an estimate of the extent to which a particular entity (detailed geographical unit or industry) differs from the average. That deviation is modeled as the realization of a random process, and is estimated conditional on the actual values of a few entities with the particular level of the detailed factor under study. To further illustrate the types of models covered by our analysis, consider the example of county job creation rates. The geographic indicator for a county in the United States is an example of a factor that has many levels (over 3,000). Estimating the difference in job creation rates for a single county, as compared to the entire country, is an example of small-area estimation. Typically, only a few businesses (a sample of those located in the county of interest) provide direct data on the level of the county job creation effect. Linear mixed-effect models combine the data from the businesses located in the county of interest with data from businesses in other counties. This statistical use of the data from other counties improves the utility of the estimated county effects and provides the potential to protect the privacy of the data from businesses located in the county of interest because such data are not the exclusive inputs to the estimated county effect.

2 Data Sources

We use the Census Bureau's Quarterly Workforce Indicators (QWI) as our application of LMM estimation to small area and industrial detail data protection. (See [1] and [2] for details on the data creation.) The QWI data contain employment counts, accessions, separations, job flows, earnings, and explanatory variables of interest—namely, industry, county (within state), and date (1990:Q2 to 2010:Q1). The dependent variables of interest here are the job creation rate (JCR), job destruction rate (JDR), accession rate (AR), and separation rate (SR). We model rates instead of levels because the differentially private estimators we consider require a bounded parameter space, and these rates are naturally bounded, which effectively bounds the parameter space for LMMs. Industry and county are categorical variables. Time is an integer-valued variable measured in quarters.

The four rates are defined in the establishment-level micro-data as $AR = \frac{A}{\bar{E}}$, $JCR \equiv \max\left(0, \frac{E-B}{\bar{E}}\right)$, $JDR \equiv \max(0, \frac{B-E}{\bar{E}})$, and $SR \equiv \frac{S}{\bar{E}}$, where $\bar{E} \equiv \frac{E+B}{2}$, E is end-of-quarter employment, B is beginning-of-quarter employment, A is accessions within the quarter, and S is separations within quarter. The identity $JCR - JDR = AR - SR$ holds for all entities and time periods [1]. In this paper we use detailed aggregates published from the micro-data. We are thus treating the detailed publication data as proxies for the micro-data as an experiment in statistical disclosure limitation methods. Only JCR, JDR, and AR are modeled since SR can be calculated from the identity. Note that JCR and $JDR \in (0, 2)$ but AR, $SR \in (0, \infty)$ by assuming $B, E, A > 0$. A value of $JCR = 2$ indicates that all jobs are born in a quarter and a value of $JCR = 0$ indicates that no jobs are born. AR and SR are empirically not very large unless an establishment j hires or separates many more employees in a quarter than it has at the beginning and end of the quarter. We use an empirical range for AR and SR because if we used ∞ for the private models, all results would result in pure statistical noise. The dependent variables of interest are specified as rates in the LMM specified in Section 3 and modeled accordingly. Categorical variables take the values of 0 or 1 in the X or Z design matrices defined below and their respective fixed effects and random effects, β and u, are therefore bounded by the range of the dependent variable.

3 Model Specifications

3.1 Linear Mixed Model

3.1.1 Background

One purpose of the classical Linear Model (LM) is to estimate the numerical relations between dependent and independent variables. The two requirements of the LM are: first, that the average value of the dependent variable, JCR, is a linear combination of known data (e.g., industry and time) and other unknown constants (β, the fixed effects); and second, that the dependent variable is normally distributed with a mean at the value of the linear combination. When some of the parameters of the LM are

treated as realizations of random variables instead of unknown constants, the model is called a Linear Mixed Model (LMM). In other words, when the mean of JCR is a linear combination of constant terms and random terms which are not constants, the model is a LMM [17]. In our case, some of the random variables are county random effects and we assume that they come from a random sample of counties from the entire population of counties. The LMM allows for JCR to depend on the county and we expect each county's random effect to be 0 with uncertainty according to a normal distribution. However, for this application we are not only interested in estimating the variance of county effects (i.e., how much JCR varies due to particular counties alone), but also in the particular level of the realized county random effect, $\hat{u}_c$. First, we define several technical definitions used in this paper in Table 1 [24].

Word	Definition
factor	the classifications (industry, county)
levels	the individual classes of a classification (manufacturing industry, construction industry)
cells	intersection of one level of every factor (manufacturing in Orange County)
balanced data	when each cell contains the same number of observations
effect	extent to which different levels of a factor affect the variable of interest
fixed effects	effects attributable to a finite set of levels of a factor that occur in the data
random effects	effects attributable to an infinite set of levels of a factor, of which only a random sample occur in the data
variance components	random effect variance and error variance
detailed factor	a factor with many levels
industrial detailed data	subindustry
design matrix	a matrix indicating which observations belong to which levels
burn-in	the number of MCMC iterations to initialize Bayesian estimation and later discard

Table 1: Technical Definitions

3.1.2 Linear Mixed Model Specification

Our statistical model is specified as follows:

$$y = X\beta + Zu + \xi, \tag{1}$$

where y $(N \times 1)$ consists of elements y_{jct}, the value of one of the dependent variables $(JCR,\ JDR,\ AR)$. The subscript j is industry (20), c is a unique county within a state $(3, 111)$, t is time (quarters from 1990:2 to 2010:1), and N is the total number of observations. X is the $(N \times 21)$ design matrix for the fixed effects (industry and the time trend). Z is the $(N \times 3, 111)$ design matrix for the random county effects, where each row has a 1 in the column of that observation's county. Finally, ξ is the $(N \times 1)$ observational random effect across all observations and is assumed independent and identically distributed. $\hat{\beta}$ is the vector of maximum likelihood estimates (MLEs) and $\hat{u}$ is the vector of empirical best linear unbiased predictors (EBLUPs). Random effects are assumed independent with a constant variance for each county and observation.

The mixed-effect likelihood function is constructed by assuming

$$\xi \sim N(0, \sigma_\xi^2 I_N) = N(0, R)$$

$$u \sim N(0, G),$$

where $R = \sigma_\xi^2 I_N$ and $G = \sigma_c^2 I_{3111}$. These assumptions imply

$$E\left[y|X, Z\right] = X\beta$$

$$y \sim N(X\beta, ZGZ^T + R) = N(X\beta, V),$$

and given random effects due to state and county

$$E\left[y|X, Z, u\right] = X\beta + Zu$$

$$(y|u) \sim N(X\beta + Zu, \sigma_\xi^2 I_N),$$

which implies Equation 1.

Henderson et al. [15] show that maximizing the joint density of y and u yields the MLEs $\hat{\beta}$ and EBLUPs $\hat{u}$ that solve:

$$X^T R^{-1} X\hat{\beta} + X^T R^{-1} Z\hat{u} = X^T R^{-1} y$$

$$Z^T R^{-1} X\hat{\beta} + Z^T R^{-1} Z\hat{u} + G^{-1}\hat{u} = Z^T R^{-1} y.$$

Additionally, we are interested in estimating the two variances, σ_ξ^2 and σ_c^2, for statistical inference and the generation of the EBLUPs.

3.1.3 Maximum Likelihood and Restricted Maximum Likelihood Estimates

To calculate all estimates of interest, we use the `lmer()` function from the R package `lme4`, which maximizes the restricted log-likelihood, called REML [5], and takes advantage of sparse matrix computational methods [3]. Table 2 shows a summary of the REML estimates produced for our model. Initial global estimates are calculated from Table 2 independently for each of the three modeled rates (JCR, JDR, and AR) using the original data (about 2.4 million rows). These estimates ($\hat{\beta}^{global}$,$\hat{u}^{global}$,$\hat{\sigma}^{global}$) act as a benchmark for the differentially private methods in this paper that use sub-sampling and Laplace noise ($\hat{\beta}^{DP_\epsilon}$,$\hat{u}^{DP_\epsilon}$,$\hat{\sigma}^{DP_\epsilon}$). The goodness of fit for the benchmark estimates is the correlation of the true rates, y, to the fitted values, $X\hat{\beta}^{global} + Z\hat{u}^{global}$, of the global model. This benchmark correlation will be compared to the correlations of the true rates, y, to the fitted values, $X\hat{\beta}^{DP_\epsilon} + Z\hat{u}^{DP_\epsilon}$, of the differentially private methods. Sections 4 and 5 provide more details.

Estimates of the LMM parameters are produced by minimizing the negative log-likelihood (MLE) or restricted log-likelihood (REML). Although there is no closed form solution for the MLE or REML of the complete parameter vector $\left(\beta, G/\sigma_\xi^2, \sigma_\xi^2\right)$ [4], [8], Bates and Debroy [4] show that intermediate REML calculations for the parameters in G/σ_ξ^2 can be expressed using a profiled log-restricted likelihood that only depends on a G/σ_ξ^2 and not $\left(\beta, \sigma_\xi^2\right)$.

Estimate	Dimension	Description
$\hat{\beta}_1$,...,$\hat{\beta}_{20}$	20	Industry (n) MLEs
$\hat{\beta}_{21}$	1	Quarter (t) MLE
$\hat{u}_1$,...,$\hat{u}_{3111}$	3,111	County (c) BLUPs
$\hat{\sigma}_\xi^2$	1	Residual Variance
$\hat{\sigma}_c^2$	1	County Variance

Table 2: Estimate Descriptions

3.2 Bayesian Linear Mixed Model

3.2.1 Background

Bayesian estimation of the LMM permits us to incorporate both *a priori* knowledge of the parameters, $\beta, \sigma_c^2, \sigma_\xi^2$, and information from the data, (y, X, Z), into the fitting of the BLMM to generate samples from the posterior distribution of $\beta, \sigma_c^2, \sigma_\xi^2$, and u. We set the prior distribution of β, σ_c^2, and σ_ξ^2 to their feasible ranges and use the samples from the posterior distributions to directly analyze the privacy properties of the fixed effects, variance components, and estimated random effects. We compare the posterior draws of the sensitive county random effects vector, u, from a BLMM fit with all observations (benchmark model) to BLMMs fit by deleting one influential

observation at a time. We then calculate the maximal differential privacy risk over all the single-row deletion experiments. This procedure produces an empirical DP_ϵ. For comparison, the variation for the benchmark model is established by comparing the posterior draws of a BLMM fit with all observations to those of a duplicated BLMM to produce an empirical ϵ due to natural variation alone. Empirical DP_ϵ equates the empirical privacy level, $\epsilon = \max(|\ln M_1|, |\ln M_2|)$, where M_1 and M_2 are the posterior odds ratios of the benchmark model and the comparison model. Results and further explanations are found in Section 5.

3.2.2 Bayesian Linear Mixed Model Specification

Our BLMM model is specified as follows:

$$y = X\beta + Zu + \xi$$

$$\xi \sim N(0, \sigma_\xi^2 I_N) = N(0, R)$$

$$u \sim N(0, \sigma_c^2 I_{3111}) = N(0, G)$$

$$\begin{aligned} \sigma_\xi^2 &\sim IW(V, \nu) \\ \sigma_c^2 &\sim IW(V, \nu) \\ \beta &\sim MVN(\mu, \Sigma), \end{aligned}$$

where $R = \sigma_\xi^2 I_N$, $G = \sigma_c^2 I_{3111}$, and y $(N \times 1)$ consists of elements y_{jct}, the value of one of the dependent variables (JCR, JDR, AR). The subscript j is industry (20), c is unique county $(3, 111)$, t is lagged quarterly rates (4×1) , and N is the total number of observations. X is the $(N \times 24)$ design matrix for the fixed effects (industry and lagged rates). Z is the $(N \times 3,111)$ design matrix for the random effects county.

The prior distributions of the variance components are multivariate Inverse-Wishart distributions (V, ν) that reduce to Inverse-Gamma distributions when V is 1. Some advantages of the Inverse-Wishart and Inverse-Gamma distributions are that their random variables are always real-valued positive definite matrices and positive reals, respectively, and they are the conjugate prior distributions for the multivariate normal and univariate normal distributions, respectively [12]. The prior distribution of the fixed effects is a multivariate normal distribution (μ, Σ) that allows for more complex covariance structures. Our model is similar to the LMM except for the additional covariates that model the time structure more accurately and the use of Markov Chain Monte Carlo (MCMC) sampling instead of REML estimation. MCMC sampling does not sub-sample observations as in the sub-sample and aggregate approach. Instead, it samples likely values of the parameter estimates using the posterior distributions. We use a high number of these parameter samples to analyze privacy implications, but desire to eventually

release only one estimate. The intermediate outputs of MCMC are draws from the posterior distribution of the parameters and the random effects while the outputs of REML are point estimates. MCMC sampling from BLMMs gives us greater flexibility in analyzing the tails of the posterior distribution of the parameters and random effects for differential privacy applications.

3.2.3 Posterior Distribution

Ten thousand samples of the parameters, $\beta, \sigma_c^2, \sigma_\xi^2$, and u, are drawn from their posterior distributions after burn-in. Then, posterior samples from the distribution of u_c (i.e., a single element of u in county c) are generated from $p(u_c|y, X, Z, \beta, \sigma_\xi^2, \sigma_c^2)$ for every county c. Section 5 has further details.

4 Differentially Private Estimation via Sub-sampling

We use LMMs and Smith's [25] differential privacy via random sub-sampling method to compute a differentially private MLE from our data by means of partitioning the complete sample into thousands of disjoint LMMs that share the same parameter vector and random effects, although only a subset of the random effects appear in any given sub-sample. The QWI data are used to form the matrices X and Z, and the vector y, which we use in this algorithm. Although we are using public data, the exercise nicely simulates protecting the confidential entity data since we are trying to summarize the characteristics of a large number of states, counties, industries, and time periods. We have not yet focused on the time effects because we are concerned with showing the effects of SDL or PPD on the small area estimates (counties within state). The time effects are given further consideration in Section 6. We apply Smith's method of differential privacy via sub-sampling [18], [25] directly to the full data matrix from the QWI.

4.1 Sub-sampling

Divide the input (y, X, Z) into k disjoint blocks, i.e., construct sub-samples by rows, $B_1, ..., B_{(i)}, ..., B_k$ of $n_k = \lfloor \frac{N}{k} \rfloor$ points each where $B_{(i)}$ denotes the i^{th} disjoint subset and N is the total number of observations. The complete data set for each of the models is denoted by $(y, X, Z) = \bigcup (y_1, X_1, Z_1)$, ..., $(y_{(i)}, X_{(i)}, Z_{(i)})$, ..., (y_n, X_n, Z_n). Using `lmer()`, calculate k sets of estimates from Table 2 using the data for each block only.

4.2 Bias-corrected $\hat{\beta}_{(i)}$ and $\hat{u}_{(i)}$

McCulloch and Searle [17] note that the solutions to Henderson's equations are $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ and $BLUP = GZ^T V^{-1}(y - X\hat{\beta})$, where $V = ZGZ^T + R$. Both of these equations are functions of at least one variance component of the model within R or G, which are not known, and must be estimated. Since the variance components

are estimated, our estimate of u is $EBLUP(u) = \hat{u} = \hat{G}Z^T\hat{V}^{-1}(y - X\hat{\beta})$. Prasad and Rao [21] state that the resulting two-stage estimator is unbiased if the expectation of the estimator is finite, the elements of the estimated variance components are even functions of y and translation-invariant, and the distributions of u and ξ are both symmetric. Our empirical results suggest that our EBLUPs are more biased as we increase the number of sub-samples k. Additionally, the estimated variance components become larger as k increases. We implemented a bias-corrected version $\hat{\beta}_{(i)}$ and $\hat{u}_{(i)}$ for the differentially private estimate generation routine. They are produced in Algorithm 1.

Data: Vectors $\hat{\beta}^{global}$, $\hat{u}^{global}$ and point estimates $\hat{\sigma}_{\xi}^{2^{global}}$, and $\hat{\sigma}_{c}^{2^{global}}$ when $k = 1$
Coefficient matrices $\hat{\beta}$ and $\hat{u}$ of dimension k by their respective number of levels
Variance vectors $\hat{\sigma}_{\xi}^{2bc}$ and $\hat{\sigma}_{c}^{2bc}$ of length k

Result: Bias-Corrected Estimates of $\hat{u}$ and $\hat{\beta}$

for $i = 1 \rightarrow k$ **do**

Compute centered vectors

$\hat{\beta}_{(i)}^{bc} = \hat{\beta}_{(i)} - \hat{\beta}^{global}$

$\hat{u}_{(i)}^{bc} = \hat{u}_{(i)} - \hat{u}^{global}$

Compute centered point estimates

$\hat{\sigma}_{\xi}^{2bc}{}_{(i)} = \hat{\sigma}_{\xi\,(i)}^{2} - \hat{\sigma}_{\xi}^{2^{global}}$

$\hat{\sigma}_{c}^{2bc}{}_{(i)} = \hat{\sigma}_{c\,(i)}^{2} - \hat{\sigma}_{c}^{2^{global}}$

$\hat{\sigma}_{(i)}^{2bc} \stackrel{d}{=} (\hat{\sigma}_{\xi}^{2bc}{}_{(i)}, \hat{\sigma}_{c}^{2bc}{}_{(i)})$

end

forall *columns* (j) *of* $\hat{\beta}$ **do**

Solve the regression equation for $\hat{\gamma}_1$ (2 X 1) and produce bias-corrected vectors $\hat{u}_{(j)}^{bc*}$

1. $\hat{\beta}_{(j)}^{bc} = \gamma_1 \hat{\sigma}^{2bc} + e_1$ where $e_1 \sim N(0, \sigma_1^2)$
2. $\hat{\gamma}_1 = ((\hat{\sigma}^{2bc})^T (\hat{\sigma}^{2bc}))^{-1} (\hat{\sigma}^{2bc})^T \hat{\beta}_{(j)}^{bc}$
3. $\hat{u}_{(j)}^{bc*} = \hat{u}_{(j)} - \hat{\sigma}^{2bc} \hat{\gamma}_1$

end

forall *columns* (j) *of* $\hat{u}$ **do**

Solve the regression equation for $\hat{\gamma}_2$ (2 X 1) and produce bias-corrected vectors $\hat{\beta}_{(j)}^{bc*}$

1. $\hat{u}_{(j)}^{bc} = \gamma_2 \hat{\sigma}^{2bc} + e_2$ where $e_2 \sim N(0, \sigma_2^2)$
2. $\hat{\gamma}_2 = ((\hat{\sigma}^{2bc})^T (\hat{\sigma}^{2bc}))^{-1} (\hat{\sigma}^{2bc})^T \hat{u}_{(j)}^{bc}$
3. $\hat{\beta}_{(j)}^{bc*} = \hat{\beta}_{(j)} - \hat{\sigma}^{2bc} \hat{\gamma}_2$

end

Relabel all values of matrices $\hat{u}^{bc*}$ and $\hat{\beta}^{bc*}$ as $\hat{u}$ and $\hat{\beta}$

Algorithm 1: Bias-Corrected Estimates

4.3 Averaging Sub-samples as the Aggregation Function

Average the estimates over k blocks:

$$\begin{aligned}\hat{\beta}^{**} &= \frac{\sum_{i=1}^{k} \hat{\beta}_{(i)}}{k} \\ \hat{u}^{**} &= \frac{\sum_{i=1}^{k} \hat{u}_{(i)}}{k},\end{aligned}$$

$$\hat{\sigma}_c^{2**} = \frac{\sum_{i=1}^{k} \hat{\sigma}_{c\,(i)}^2}{k}$$

and

$$\hat{\sigma}_\xi^{2**} = \frac{\sum_{i=1}^{k} \hat{\sigma}_{\xi\,(i)}^2}{k}.$$

Next, draw R_β^ϵ, R_u^ϵ, and R_σ^ϵ from independent Laplace distributions, as a function of the differential privacy parameter ϵ, where the Laplace scale parameters $b = (b_1, b_2, b_3)$ are $\frac{\Lambda_\beta}{k\epsilon}$, $\frac{\Lambda_u}{k\epsilon}$, and $\frac{\Lambda_\sigma}{k\epsilon}$, respectively, and $\hat{\sigma} = \left(\sqrt{\hat{\sigma}_c^{2**}}, \sqrt{\hat{\sigma}_\xi^{2**}}\right)$. The values Λ_β, Λ_u, and Λ_σ are the global sensitivities [25] or the maximum ranges of the parameters β, μ, and σ, respectively, as shown in Table **??**. Output $\hat{\beta}^{DP_\epsilon} = R_\beta^\epsilon + \hat{\beta}^{**}$, $\hat{u}^{DP_\epsilon} = R_u^\epsilon + \hat{u}^{**}$, and $\hat{\sigma}_\xi^{DP_\epsilon} = R_\sigma^\epsilon + \hat{\sigma}^{**}$ as the differentially private estimates with protection ϵ.

In the process of sub-sampling disjoint subsets from over 2.4 million observations or rows in the matrices X and Z, individual subsets could contain between 271 (for $\epsilon = 1$) and 500 (for $\epsilon = 4.6$) observations where the sample size of the individual subset is $n_k = \lfloor \frac{N}{k} \rfloor = \left\lfloor \left(\frac{N\epsilon}{\Lambda}\right)^{2/5} \right\rfloor$ as derived in Section 4.4. For large values of k, it is very likely that many of the sub-samples do not have entries for some industries or thousands of counties in the $X_{(i)}$or $Z_{(i)}$ matrices due to chance or the limited number of rows (n_k). Consequently, many of their respective parameters cannot not be estimated. In such cases, we treat these non-estimable $\hat{\beta}_{(i)}$ and $\hat{u}_{(i)}$ as not relevant, and do not use them in our averaged estimates. In cases with even smaller individual subsets (e.g., when $k = 16,000$, $n_k = 151$), it is possible that the mixed model is not estimable.[1] Therefore, we must keep k at a reasonable level, and in Section 5, we use $k = 8,945$ ($\epsilon = 1$) through $k = 4,858$ ($\epsilon = 4.6$).

County effects were considered random since there were 3,111 unique counties and each sub-sample contained a random subset of counties. For $k = 8,945$ in all sub-samples, industry had at most 1 or 2 industries missing from X and county had between

[1] Whenever we say "not estimable," we mean that the relevant moment matrix is singular. In practice, this means that the linear mixed-effects model must be reduced in dimension before any of the levels of the effects of interest can be estimated. Rather than impose arbitrary dimension reductions, we labeled such models "not estimable."

2,863 and 2,892 counties missing from Z. In cases where there were many unique counties missing, the estimated variance of the random effect due to unique county, $\hat{\sigma}_c^2$, was often zero due to the lack of repeated observations per unique county. Low prevalence categories in industries, such as industry 92 (Public Administration), had many fewer observations than others and, consequently, had higher ranges of estimated coefficients from sub-sample to sub-sample.

The Laplace scale parameters, b, are dependent on Λ, ϵ, and k. By fixing k and Λ, the resulting Laplace scale parameters become a function of ϵ alone, which we tried to vary from 0.1 to 4.6, but models using values less than unity for ϵ were not estimable.

The Λ_β, Λ_u, and Λ_σ are maximum ranges in the corresponding parameters of β, u, and σ, as shown in Table **??**. For JCR and JDR, all three components of $\hat{\sigma}$ are bounded since standard deviations should be a maximum of 0.5 for rates $\in (0, 2)$. So, we set Λ_σ to 0.5. $\hat{\beta}$ and $\hat{u}$ depend on the scale of the data, which in our case contains only 0 and 1 except for the time trend variable. In such binary cases and disregarding any interactions, we set Λ_β and Λ_u to be 2. In simulations, the quarter estimate, $\hat{\beta}_{21}$, always had a very low range across sub-samples (less than .004), so $\Lambda_{\beta_{21}}$was set to .01 because 2 would be too large for the scale of the time trend variable (values of 22 to 101).

For AR, the bounds need to be larger to account for the empirical range of AR $\in (0, 385)$, which is much too large for meaningful statistical inference when this range is used to set the Laplace scale parameter for the differentially private estimates. The theoretical bound would be ∞, which would render all analysis unusable. Therefore, we make a relaxation for our analysis and note that data collection without theoretical bounds is not likely to be differentially private unless future data collection efforts are modified. We calculate empirical ranges of the parameter estimates over different values of k for all rates in Table **??**. When looking at the 0.1% to 99.9% quantile of AR, the rates are $\in (0, 2.57)$. Consequently, $\Lambda_\sigma$was set to 0.75, Λ_β and $\Lambda_u$were set to 3, and $\Lambda_{21}$was set to 0.01 (from empirical simulation). Table **??** shows the maximum ranges of estimates for JCR and AR over $k = 8,945$ sub-samples, which always had larger maximum ranges than smaller k in our simulations. Due to the nature of how the rates were measured [1], the maximum range of JCR is theoretically and empirically at most 2, however, the theoretical range of AR is unbounded which is why it was limited at the 99.9% empirical quantile.

4.4 Number of Sub-samples

Smith [25] shows that the maximum number of sub-samples to be considered is $k = n^{2/3}$ to get a sufficiently small bias, and the optimal number of sub-samples is $k^* = \frac{n^{3/5}\Lambda^{2/5}}{\epsilon^{2/5}}$ to get an asymptotic relative error that tends to 1. Setting Λ_β and Λ_u equal to the maximum of all estimate ranges for the JCR and JDR models implies an optimal k^* of $\frac{8941}{\epsilon}$. As ϵ ranges from 0.1 to 4.6, the optimal k^* ranges from 22,470 to 4,858. Results are presented using $\epsilon \in (1, 2, 3, 4, 4.6)$. A value of $k^* > 9,000$ is not feasible

within the REML computation because the low sample size ($n_k = 151$) does not permit any estimation at all. Other values of k can be considered and produce the following equivalence table in 1 for $\Lambda = 2$ and $N = 2,428,452$. The number of sub-samples required for the empirical range of AR would be more than eight times that of Figure 1.

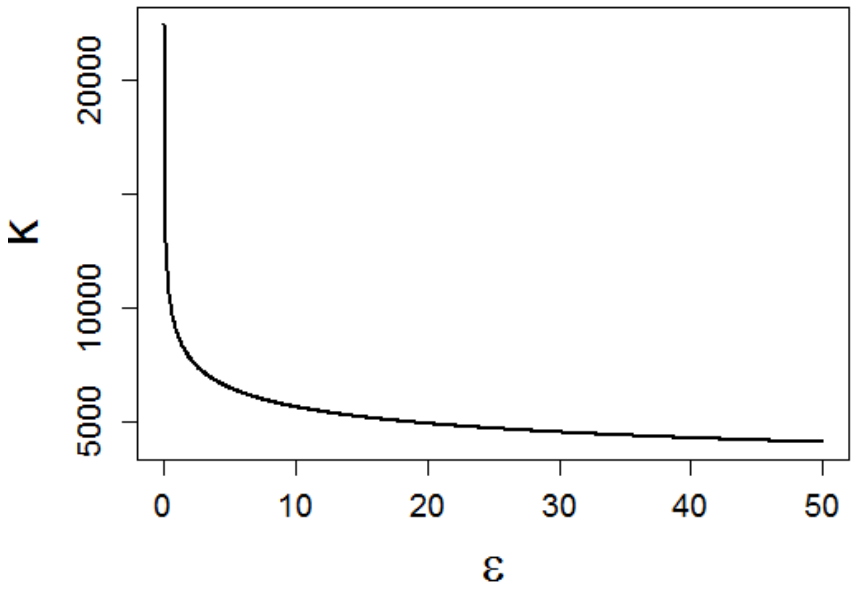

Figure 1: Equivalence table for optimal k over values of ϵ for JCR

4.5 Differentially Private Fitted Values

The fitted values of our mixed model are linear combinations of the rows of X and Z and the differentially private estimates with protection ϵ. X and Z are sparse matrices because the columns are categorical variables and any given row is identified by an industry, unique county, and quarter. Any fitted value is the sum of three differentially private estimates with protection ϵ and a quarter, t, times the differentially private trend estimate, $\beta_{21}^{DP_\epsilon}$, with protection ϵ. Ignoring the differentially private trend estimate, $\beta_{21}^{DP_\epsilon}$, and assuming each row can only change by industry and unique county, we provide a proof for differentially private fitted values that builds on Smith's proof [25].[2]

[2] $t\Lambda_{\beta_{21}} \leq 101(.01) = 1.01$. So, $t\Lambda_{\beta_{21}} < \Lambda_\beta$ and by using the Laplace scale parameter, R_β^ϵ, for β_{21}, $\beta_{21}^{DP_\epsilon}$ is also ϵ-differentially private.

Lemma 1. *For any choice of the number of sub-samples k, a fitted value for any row is $C\epsilon$-differentially private where C is 2, the assumed number of non-zero entries in X and Z for an added or deleted row r.*

Proof. Given fixed matrices of X and Z, consider adding or deleting a row r to obtain neighbor matrices X' and Z' that differ from X and Z by only by one observation or row. At most, only one of the sub-samples $B_{(i)} = (y_{(i)}, X_{(i)}, Z_{(i)})$ can include or exclude row r. The maximum that the components of $\hat{\beta}_{(i)}$ and $\hat{u}_{(i)}$ can change with or without row r is by their global sensitivities, Λ_β and Λ_u. Therefore, the most $\hat{\beta}^{**} = \frac{\sum_{i=1}^{k} \hat{\beta}_{(i)}}{k}$ and $\hat{u}^{**} = \frac{\sum_{i=1}^{k} \hat{u}_{(i)}}{k}$ can change are $\frac{\Lambda_\beta}{k}$ and $\frac{\Lambda_u}{k}$, respectively. By Smith [25], this results in the Laplace random variables $\hat{\beta}^{DP_\epsilon} = R^\epsilon_\beta + \hat{\beta}^{**}$ and $\hat{u}^{DP_\epsilon} = R^\epsilon_u + \hat{u}^{**}$ each being ϵ-differentially private where the Laplace noises were defined in Section 4.3. Define an arbitrary fitted value of the vector $\hat{y}^{DP_{C\epsilon}} = X\hat{\beta}^{DP_\epsilon} + Z\hat{u}^{DP_\epsilon}$ as $\hat{y}^{DP_{c\epsilon}}_a$. $\hat{y}^{DP_{c\epsilon}}_a$ is a function of two ϵ-differentially private estimators without using additional confidential data (X and Z are not confidential) and therefore, 2ϵ-differentially private. □

Note that the proof can be generalized to different allocations of privacy, such as two estimators that are 0.1ϵ-differentially private and 0.9ϵ-differentially private by changing the Laplace scale parameters. The result is that a fitted value for any row would be $(0.1 + 0.9)\epsilon$-differentially private or ϵ-differentially private. We generalize the proof to three differentially private estimators for the industry effect, trend effect, and unique county EBLUP. All figures use a total of ϵ-differential privacy with varying levels of the privacy budget for β and u, and an allocation of 2% of ϵ for β_{21}. Additionally, we did 30 random simulations of differentially private fitted values and averaged the correlation results in Figures 3, 4, 5, 6, 7, and 8.

5 Differentially Private Estimation via Expected Risk Minimization

We use BLMMs for Chaudhuri, Monteleoni, and Sarwate's [6] approach of differential privacy and relate the posterior distribution of the BLMM to ERM. Their approach shows that ϵ-differential privacy can be obtained by perturbing an objective function, J_{priv}, to obtain an efficient, differentially private approximation for the predictors, $\mathbf{f}_{priv}$, of regularized ERM.

$$\begin{aligned} \mathbf{f}_{priv} &= \arg\min J_{priv}\left(\boldsymbol{f}, \mathcal{D}\right) + \frac{1}{2}\Delta \left\|\mathbf{f}\right\|^2 \\ &= \arg\min[\frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathbf{f}\left(\mathbf{x}_i\right), \mathbf{y}_i\right) + \Lambda N\left(\mathbf{f}\right) + \frac{1}{n}\boldsymbol{b}^T\mathbf{f}] + \frac{1}{2}\Delta\left\|\mathbf{f}\right\|^2. \end{aligned}$$

$\mathbf{f}_{priv}$, or more commonly known as regression coefficients (β), is obtained by minimizing a loss function and a regularizer [6]. One major difference between their approach and ours

is that the objective perturbation algorithm relies on classifiers for binary dependent data and our application has continuous, bounded dependent variables. The original Chaudhuri et al. algorithm shows that global sensitivity comes from the assumption that the loss function is convex and bounded, has a strictly convex penalty term, and has a smooth and bounded derivative. In our application, we use bounded continuous rates and define an informative prior distribution that bounds the parameters in the posterior distribution from which we calculate the empirical level of ϵ.

We note that regularized risk minimization is equivalent to maximum a posteriori estimation and

$$\begin{aligned}
\arg\min[\frac{1}{n}\sum_{i=1}^{n}\ell\left(\mathbf{f}\left(\mathbf{x}_i\right),\mathbf{y}_i\right)+\Lambda N\left(\mathbf{f}\right)] &= \arg\min[\text{Empirical Risk}+\text{Regularizer}] \\
&= \arg\min[-\log L()-\log p(\mathbf{f})] \\
&= \arg\max\left[\log\left(L()\times p(\mathbf{f})\right)\right] \\
&= \arg\max\left[L()\times p(\mathbf{f})\right] \\
&= \arg\max\left[\text{posterior}\right].
\end{aligned}$$

We proceed in the Bayesian fashion by setting priors on our fixed effects and variance components. Then, we fit the complete-data model. Next we remove influential observations one at a time in order to estimate the effective ϵ-differential privacy of the complete-data procedure. We analyze effects on the posterior distribution of the complete set of u_c, the county random effects, because these are much more sensitive to a breach in privacy than the fixed effects or variance components.

5.1 Prior Specification

We use the Inverse-Wishart distribution $IW(V,\nu)$ and multivariate normal distribution $MVN(\mu,\Sigma)$. To test the feasibility of differential privacy for the BLMM we consider the simplest case of such distributions. Hence, we use the univariate Inverse-Wishart distribution with mean $\frac{\nu V}{\nu-2}$ and variance $\frac{\nu^2V^2}{(\nu-2)^2\left(\frac{\nu}{2}-2\right)}$ for the two random effects. We also set the prior of our fixed effects at $\mu=0$ and $\Sigma\propto I$ causing the multivariate normal distribution to produce independently and identically distributed univariate normal distributions with mean zero and constant variance a priori. In order for our priors to depend on only one parameter, ν, the degrees of freedom, we set V equal to a constant. Our benchmark and confidential prior distribution, p_0, is diffuse with the bounds being set as close as possible to the feasible ranges of the parameters $\beta\in(-2,2)$ and σ_c^2, $\sigma_\xi^2\in(0,0.25)$. When setting $V=0.104$ and $\nu=12$, the prior mean and standard deviation of σ_c^2 and σ_ξ^2 are 0.125 and 0.0625, which we define as the benchmark prior, p_0, that spans the feasible range of our variance components. We also set $\Sigma=16^2\frac{v^2V^2}{(\nu-2)^2\left(\frac{\nu}{2}-2\right)}I$ to ensure the standard deviations of our benchmark univariate normal priors are 1, span the feasible range of β, and scale with the priors of σ_c^2 and σ_ξ^2. This gives the following BLMM

$$
\begin{aligned}
Y &= X\beta + Zu + \xi \\
R &= \sigma_\xi^2 I \\
G &= \sigma_c^2 I
\end{aligned}
$$

$$
\begin{aligned}
p\left(\sigma_\xi^2|V,\nu\right) &= \frac{|\nu V|^{\frac{\nu}{2}}}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}|\sigma_\xi^2|^{-\frac{\nu+2}{2}}\exp\left(-\frac{1}{2}\frac{\nu V}{\sigma_\xi^2}\right) \\
p\left(\sigma_c^2|V,\nu\right) &= \frac{|\nu V|^{\frac{\nu}{2}}}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}|\sigma_c^2|^{-\frac{\nu+2}{2}}\exp\left(-\frac{1}{2}\frac{\nu V}{\sigma_c^2}\right) \\
p(\beta|\Sigma) &= (2\pi)^{-d/2}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\beta^T\Sigma^{-1}\beta\right)
\end{aligned}
$$

where $\nu > 0$, $V > 0$, d is the dimension of β, and $\Sigma = 16^2\frac{v^2V^2}{(\nu-2)^2\left(\frac{\nu}{2}-2\right)}I$.

5.2 Bayesian Computation and ϵ-Differential Privacy

We use the R package `MCMCglmm` to fit the BLMM through MCMC simulation. `MCMCglmm` uses C++, samples all location parameters in a single block, uses CSparse C libraries, and is 40 times faster than Winbugs [13]. Even with these advantages, for our data it takes about 10 hours to run 20,000 MCMC iterations with a 10,000 iteration burn-in and thinning interval of 5. To incorporate the intuitive notion of differential privacy for the sensitive county random effects, we remove one observation from our data set and rerun the BLMM to generate the new posterior distribution of u. We use influence diagnostics geared for the LMM to choose those observations that require closer examination for the BLMM. MCMC methods sample from the probability distributions of the parameters and not the observations themselves as was done in the sub-sampling approach. We infer about the differential privacy properties of the model by looking at the changes in the probability distributions between the benchmark model and the model missing one influential observation.

5.2.1 Influential Observations

We delete observations i that are most influential on the EBLUPs of our LMM under REML estimation and later fit a separate BLMM for each of those observations' deletions. Traditional influence diagnostics for the LM are not completely transferable to the LMM because $\hat{\beta}$ and $\hat{u}$ are functions of the estimated variance components, σ_ξ^2 and σ_c^2. For example, the residuals of the LMM do not have to sum to zero and can sometimes produce negative values of leverage located on the diagonals of the "hat matrix," $H = X(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}$ [23]. Since the LMM should be refit when deleting each observation i to know exactly how estimates change, we incorporate several notions from the influence diagnostic literature to select observations that are influential on the entire model or specific EBLUPs.

Define the marginal residuals given the fixed effects as $y_i - x_i'\hat{\beta}$ and the conditional residuals given the EBLUPs as $r_i = y_i - x_i'\hat{\beta} - z_i'\hat{u}$. Assuming the variability of $\hat{\beta}$ is negligible given the sample size of our data, we calculate the Pearson residuals given the conditional variances as $r_i^p = \frac{r_i}{\sqrt{Var(Y_i|u)}} = \frac{r_i}{\sigma_\xi^2}$, which in our LMM is simply proportional to r_i due to the simple covariance structure and conditioning on the EBLUPs. Schabenberger suggests calculating conditional residuals and conditional Pearson residuals for influence diagnostics in the LMM [23]. Nobre and Singer [19] suggest looking at a standardized version of the conditional residuals by dividing r_i^p by a function of the joint leverage of the fixed and random effects for detecting the presence of outlying observations. They also reference Pinheiro and Bates [20], who suggest looking at extreme values of $\hat{u}$ for detecting the presence of outlying EBLUPs.

We selected 32 influential observations to examine. First, we selected 14 observations with the most extreme positive or negative r_i^p values. Second, since our design matrix, Z, is unbalanced with five counties containing fewer than 15 observations, we selected 10 observations with the minimum and maximum r_i^p from each of those counties. Finally, we selected eight observations with the minimum and maximum r_i^p from four counties with extreme values of $\hat{u}$.

5.2.2 Differential Privacy of the Realizations of County Random Effects

We develop a methodology for calculating empirical ϵ-differential privacy for continuous data using the posterior distribution of u from our BLMMs after model estimation. Because this is done after model estimation instead of before, we use the term "empirical differential privacy." First, we generate 10,000 samples from the posterior distribution of β, u, σ_c^2,and σ_ξ^2 from our benchmark model with prior density specified in Section 5.1 using all observations and after discarding the 10,000 burn-in samples. Next, we remove an influential observation i, refit our model with the same prior, and then generate 10,000 posterior samples—again, after discarding 10,000 burn-in samples. We estimate the changes for the tails of the posterior distribution of u between the benchmark model and the model deleting an influential observation.

For our models let

$$\begin{aligned} D &\equiv (y, X, Z), \text{ entire data set} \\ \text{and } D^{\tilde{}i} &\equiv (y^{\tilde{}i}, X^{\tilde{}i}, Z^{\tilde{}i}), \text{ the data set without observation } i. \end{aligned}$$

Define the posterior odds for the two data structures as $\frac{P(D^{\tilde{}i}|u_c \in b)}{P(D|u_c \in b)}$ and the prior odds as $\frac{P(D^{\tilde{}i})}{P(D)}$. For all county-effect posterior distributions of the random effect u_c, we discretize the posterior distribution of u_c to make these probability statements estimable. Let $b = 1, ..., B$ denote the bins of this discretization, where the boundaries are set such that the posterior distribution, $P(u_c \in b|D)$, estimated from the complete data has

$$P(u_c \in b|D) = \frac{1}{B},$$

where the notation $u_c \in b$ means that u_c is contained in the set whose upper and lower

bounds define the interval b in the discretization. Bounding the maximum and minimum posterior odds ratio

$$M_1 \equiv \max_{i,c,b}\left[\frac{P(u_c \in b|D^{\sim i})}{P(u_c \in b|D)}\right] = \max_{i,c,b}\left[\frac{\frac{P(D^{\sim i}|u_c \in b)}{P(D|u_c \in b)}}{\frac{P(D^{\sim i})}{P(D)}}\right]$$

and

$$M_2 \equiv \min_{i,c,b}\left[\frac{P(u_c \in b|D^{\sim i})}{P(u_c \in b|D)}\right] = \min_{i,c,b}\left[\frac{\frac{P(D^{\sim i}|u_c \in b)}{P(D|u_c \in b)}}{\frac{P(D^{\sim i})}{P(D)}}\right],$$

where the second equality is due to Bayes law. The expressions for M_1 and M_2 are equivalent to defining empirical $\epsilon-$differential privacy as $\epsilon = \max(|\ln M_1|, |\ln M_2|)$.

The empirical differential privacy measure defined here, $\epsilon = \max(|\ln M_1|, |\ln M_2|)$, means that the risk measure used for statistical disclosure limitation is the probability of an inferential disclosure as originally specified by Dalenius [7]. His definition of an inferential disclosure is the right-hand side of the definitions of M_1 and M_2. Hence, we are implementing a procedure that limits the probability of an inferential disclosure by bounding the odds ratio for such a disclosure using the differential privacy bound, the left-hand side of the definitions of M_1 and M_2. The empirical privacy level in one data set may be significantly different from the level on a neighboring data set, where one element has been deleted as we specify in our definition of $D^{\sim i}$.

One method of calculating the posterior odds is to fit a kernel density estimator of the posterior samples of u, and then evaluate these ratios over narrow bin widths. We found this method to be overly sensitive to posterior samples in the tails of the posterior distribution. Instead, we approximate $\max(|\ln M_1|, |\ln M_2|)$ by comparing the outcomes in the benchmark model, $P(u_c \in b|D)$ with outcomes in models estimated deleting an influential observation, $P\left(u_c \in b|D^{\sim i}\right)$, using a discretized posterior with 20 bins whose boundaries are determined by $P(u_c \in b|D)$.

Given 10,000 posterior samples from $u_c|D$, the benchmark model, we create 20 equal-probability bins using 500 samples corresponding to the five percent quantiles of these posterior samples. Then, for each model with deleted observation i, we count the number of posterior samples, $n_{i,c,b}$ from $u_c|D^{\sim i}$ within each of the benchmark bins. Over all models without i, county random effects ($c = 1, 2, ..., 3111$), and bins ($b = 1, 2, ..., 20$), compute $\frac{n_{i,c,b}}{500}$ and set $\epsilon = \max(|\ln M_1|, |\ln M_2|)$ where $M_1 = \max_{i,c,b}\left[\frac{n_{i,c,b}}{500}\right]$ and $M_2 = \min_{i,c,b}\left[\frac{n_{i,c,b}}{500}\right]$.

5.2.3 Convergence

We monitored the convergence of the benchmark model by performing two iterative simulations (with dispersed initial conditions) and evaluating the Gelman and Rubin convergence diagnostic. Each simulation was run for 10,000 iterations after a burn-in of 10,000 samples. The Gelman and Rubin convergence diagnostic measures the between-sequence variance, BV, and the within-sequence variance, WV, for two or more iterative

sequences. It outputs a potential scale reduction factor, $\sqrt{\frac{\frac{n-1}{n}WV+\frac{1}{n}BV}{WV}}$, that declines to 1 as the number of posterior samples, n, goes to infinity [12]. Gelman, Carlin, Stern, and Rubin note that for most examples, scale reduction factors below 1.1 are acceptable. The upper confidence limits of the potential scale reduction factors for our 3,111 county-wide random effects, two variance components, and 24 fixed effects were always between 0.99990 and 1.0047 except for county random effect u_{1460} at 1.2853 which only had 58 observations. We examined the trace plots for county random effect 1460 in Figure 2 below and found no issues with convergence.

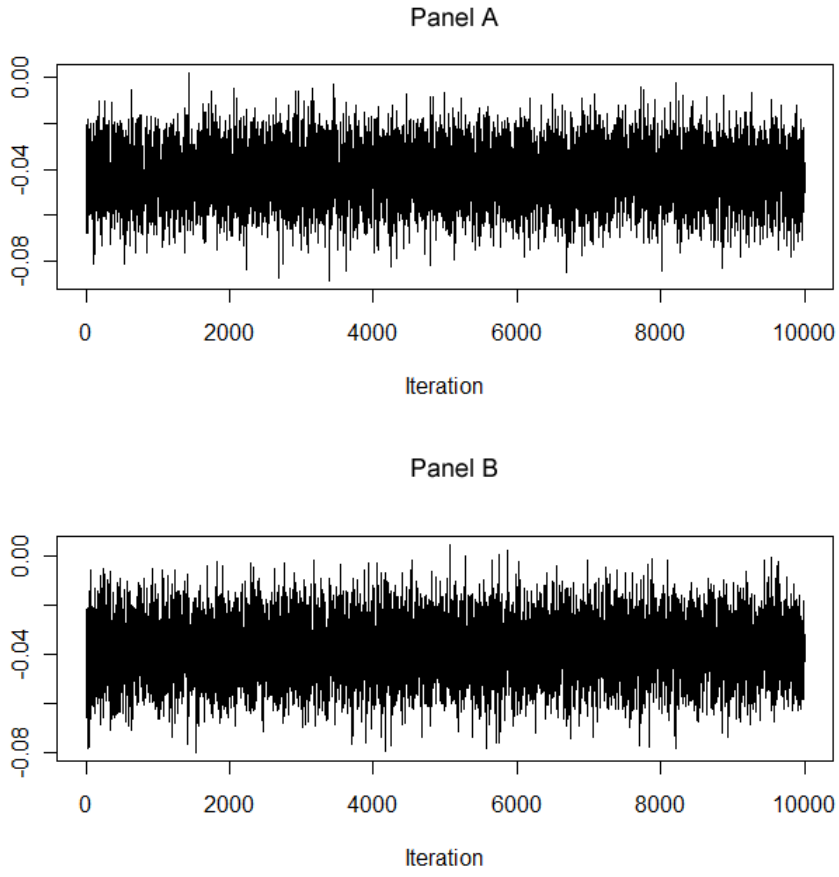

Figure 2: Trace Plots for County 1460. Panel A is the estimated random effect for 10,000 MCMC samples, after burn-in, using all data and the first set of initial conditions. Panel B is the estimated random effect for 10,000 MCMC samples, after burn-in, using all data and the second set of initial conditions.

6 Results

6.1 Linear Mixed Models

We produced $R-U$ (Risk-Utility) curves or $R-U$ confidentiality maps that examine the trade-off between ϵ (disclosure risk) and correlations (data utility) by changing parameter values in our procedure. Duncan [9] states that "in its most basic form, an $R-U$ confidentiality map is the set of paired values (R,U) of disclosure risk and data utility that correspond to various strategies for data release." In our models, ϵ changes to generate the $R-U$ curve and lower values of ϵ correspond to lower levels of risk and higher levels of privacy. As ϵ decreases, the privacy of our released data increases as defined by ϵ-differential privacy. Low disclosure risk has good differential privacy, which says that "any possible outcome of an analysis should be "almost" equally likely, independent of whether any individuals opt into or opt out of the data set" [10]. In addition, since the Laplace scale parameter is $\frac{\Lambda}{k\epsilon}$, the random noise added to release ϵ-differentially private data increases as ϵ decreases (k increases at a slower rate than ϵ decreases). This means that the released data or estimates are more noisy for lower values of ϵ. Consequently, data utility should be lower for released data with more noise added. We examined the exact trade-off between disclosure risk, ϵ, and data utility, the correlation of ($\hat{y}^{DP_\epsilon}$,y). The value on the x-axes labeled "MLE" is the non-private benchmark model and the other values are the private ϵ values increasing in privacy from 4.6 to 1.0. "MLE" is associated with an extremely high value of ϵ.

6.1.1 $R-U$ Curve for Linear Mixed Models

For all values of ϵ, calculate the predicted rates:

$$JCR^{DP} = \hat{y}^{DP_\epsilon} = X\hat{\beta}^{DP_{.51\epsilon}} + Z\hat{u}^{DP_{.49\epsilon}}.$$

For $k = 1$ or all of the data, calculate the predicted rates:

$$JCR^{global} = \hat{y}^{global} = X\hat{\beta}^{global} + Z\hat{u}^{global}.$$

Calculate the correlations between y and $\hat{y}^{global}$, $\hat{y}^{DP_\epsilon}$. Finally, plot the correlations as a function of ϵ.

6.1.2 $R-U$ Curve for Linear Models

For all values of ϵ, calculate the predicted rates:

$$JCR^{DP_\epsilon} = \hat{y}^{DP_\epsilon} = X\hat{\beta}^{DP_\epsilon}.$$

For $k = 1$ or all of the data, calculate the predicted rates:

$$JCR^{global} = \hat{y}^{global} = X\hat{\beta}^{global}.$$

Calculate the correlations between y and $\hat{y}^{global}$, $\hat{y}^{DP_\epsilon}$. Finally, plot the correlations as a function of ϵ. The LM only estimated industry means and did not include a time trend. It is considered a fixed effects model.

Figures 3 and 4 show the R-U Curves for the LMMs and LMs, respectively. Correlations decreased as ϵ decreased, and all correlations of $\hat{y}^{DP_\epsilon}$ with y were lower than the global "best fit" correlation when $k = 1$ (which would correspond to non-differentially private $\epsilon > 25$). Since all correlations including the one between y and $\hat{y}^{global}$ were less than 0.40, the model did not fit the data well. This illustrates the principle limitation of the differentially private estimator—more random effects were required to get a good fit, detailed industry and time effects in particular, but such models were only feasible when $\epsilon > 25$, which is no protection at all. But for models with approximately 3,000 effects, degradation in correlation over decreased values of ϵ was only slightly noticeable. Non-monotonicity was observed when most of the noise was added to β versus u since there were only 21 random Laplace draws.

6.1.3 R-U Curve for Linear Mixed Models with Allocated Privacy

Additionally, we considered having proportionally different levels of privacy for β and u within the total privacy budget of ϵ. Since there were many more estimates of u (3,111) as compared to industry β (20), it may be reasonable to protect the estimates of u with more privacy (lower ϵ). Figures 5 and 6 show u having 10% and 88% of the plotted value of ϵ, respectively, while β accounts for the remainder. For example, in Figure 5, the five budgets of ϵ used for u were $0.46, 0.4, 0.3, 0.2$, and 0.1, while the budgets used for β were $4.05, 3.52, 2.64, 1.76$, and 0.88. Figures 7 and 8 show u having 1% and 97% of the plotted privacy value of ϵ, respectively. Noticeable degradation is seen in Figure 7 when u is highly protected. For all figures except Figure 4, the privacy budget of the time trend was kept at 2%.

6.1.4 An Improved LMM and Influential Observations

We also examined the effects of deleting all of a county's U observations on the estimates of variance components and fixed effects with a LMM that included four parameters for lagged quarterly rates. The base fit of the model improved significantly to a correlation of 49.67% as compared to just under 35% for the LMM including only a simple time trend. The goal was also to bound the possible leave-1-out changes of our REML estimates, $\hat{\beta}, \hat{\sigma}^2_\xi$,and $\hat{\sigma}^2_c$ for closer inspection for both the LMM and BLMM. We performed over three thousand leave-U-out simulations for each county. The number of observations removed in each of the simulations ranged from 4 to 1,481 with a median of 674. For each simulation, the process is as follows:

- Define $D^{\tilde{}U} \equiv (y^{\tilde{}U}, X^{\tilde{}U}, Z^{\tilde{}U})$, differing by one county-industry combination (U-out);
- Fit REML estimates and analyze changes in $\hat{\beta}, \hat{\sigma}^2_\xi$,and $\hat{\sigma}^2_c$.

Results for the leave-U-out fixed effects indicate that all industry estimates are within 0.0002 of each other except for public administration, which has a range of 0.003. The

covariates for the lagged quarterly rates were all within 0.001 of each other. The 0.1 and 99.9 percent quantiles for the variance components are described in Table 4.

Variance Component	MLE	0.1% quantile	99.9% quantile
$\hat{\sigma}^2_\xi$	0.01045409	0.01043703	0.01046005
$\hat{\sigma}^2_c$	0.00016302	0.00015722	0.00016330

Table 4: Variance Estimate Ranges from Leaving out One County

Results from the analysis of deleting influential observations from Section 5 indicate that all updated estimates of fixed effects and variance components are well within the bounds of the leave-U-out changes. The county EBLUPs that were most affected by the removal of influential observations were always the particular counties that these observations were in. The maximum change for the EBLUPs was 0.007828 (observation from county 3047) and the industry fixed effects were 0.00008281 (observation from county 661). Both of these observations came from observations with large $\hat{u}$s and large absolute values of r_i^p. If we were to match these maximum changes corresponding to four times the standard deviation of a Laplace random variable, they produce Laplace scale parameters of 0.0015 and 0.000016, respectively. To put things in perspective, the Laplace scale parameter for the estimated fixed effects and EBLUPs in the sub-sample and aggregate approach when ϵ was unity was approximately 0.0002 and when ϵ was 3 was approximately 0.00012. With no sub-sampling and the removal of an influential observation, the Laplace noise would only protect the fixed effects. Results for the BLMM focus on the county random effects.

6.1.5 Smaller Area Interactions

Model fit improves by adding more detailed factors such as county-by-seasonal interactions, however, the differentially private MLE is not estimable for values less than 3. Figure 9 and 10 show the results of the LMM with an additional random effect, u_s, which has over 12,000 levels. Model fit improves to over 44%, but degrades more quickly with larger protections levels for the smaller levels. The improved model and updated variance components are described below:

$$y = X\beta + Zu + \xi$$

$$\xi \sim N(0, \sigma^2_\xi I_N) = N(0, R), R = \sigma^2_\xi I_N$$

$$u_s \sim N(0, \sigma^2_s I), u_c \sim N(0, \sigma^2_c I_{3111})$$

$$u = (u_s^T, u_c^T)^T \sim N(0, G)$$

$$G = \begin{bmatrix} \sigma^2_s I & 0 \\ 0 & \sigma^2_c I_{3111} \end{bmatrix}.$$

6.2 Bayesian Linear Mixed Models

We analyzed the implications of the removal of influential observations on the ϵ-differential privacy of our county random effects according to Section 5.2.2. Predictably, in those models deleting observations from small counties (3047 and 661) produced the largest proportional bin changes across all models. Each model had 62,220 bins corresponding to 20 bins for each of the 3,111 county random effects. The model deleting an observation from county 3047 had as few as 21 posterior samples in its smallest bin (3,217 in its largest) and the model deleting an observation from county 661 had 3,458 posterior samples in its largest bin (26 in its smallest). Both of these unusual counts occurred in the county effect from which the influential observation was deleted. Comparing these results to the benchmark model with 500 observations in each bin and using the methodology developed in Section 5.2.2, this corresponds to an overall ϵ of 3.2.

We compared these results to random noise which is represented by the replicated benchmark model that was fit to monitor convergence in Section 5.2.3. The bin boundaries were fixed at the five percent quantiles from the complete-data estimation. Hence, the expected count in each bin is 500. The replicated model using the complete data had its smallest bin containing 382 posterior samples and its largest bin containing 641 posterior samples. This corresponds to an overall ϵ of 0.27, which is illustrated in the histogram of the bin counts shown in Figure 11 where the mode is 500, the distribution is symmetrical, and the minimum and maximum on the horizontal axis define the inputs to computing ϵ. Since no rows have been excluded from this experiment, the interpretation of ϵ is the deviation in the empirical differential privacy that results from the imprecision of using 10,000 posterior samples.

The 32 models with deleted influential observations always had maximum and minimum bin counts between the extremes of the replicated benchmark model and the models with deleted observations from county 3047 or county 661. That is, the extreme values used to estimate ϵ empirically came from the values computed when influential observations were deleted from one of these two counties. A histogram of the bin counts for the model deleting an influential observation from county 3047, which defined the overall ϵ of 3.2, is shown in Figure 12.

7 Discussion

Results are presented for JCR only; however, JDR and AR give similar results. The main difference in the structure of AR is Λ, which is slightly larger. Thus, the Laplace scale parameter is also larger to account for the greater range of AR. In general, the more private we make our confidential data through Laplacian noise, the less information utility we receive from the released data. In this case, information utility translates to estimates of the differentially private JCRs ($\hat{y}^{DP_\epsilon}$) that are produced from differentially private coefficient estimates ($\hat{\beta}^{DP_\epsilon}$ or $\hat{u}^{DP_\epsilon}$). We note that the non-private MLE for this problem doesn't fit very well, and the differentially private MLEs are quite comparable – that is, they aren't much worse. The problem arises when we try to improve the

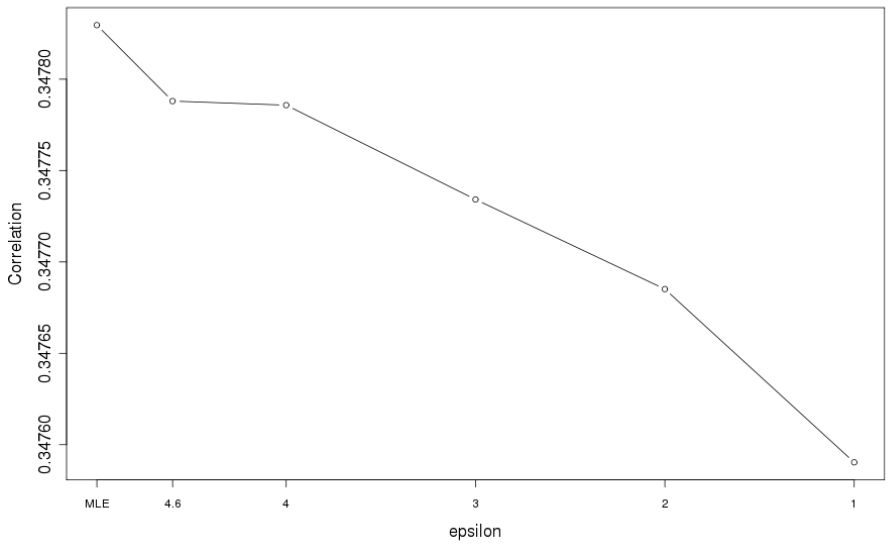

Figure 3: R-U Curve for JCR Linear Mixed Model with 49% ϵ budget for β and 49% for u

fit of the base MLE; then, we must add more effects (factors with a large number of levels) to the model and the differentially private MLE becomes infeasible. Moving from the fixed effects model of main industry effects to including county areas improved the fit from about 30% to 35%, but the differentially private MLE was not computable at values below one. After accounting for seasonal by county interactions, the base fit improves from 35% to 44%, however the differentially private MLE is not computable for privacy levels less than three. This demonstrates the trade-off between model fit and ϵ-differential privacy for the sub-sample and aggregate approach.

The empirical DP analysis based on the BLMM shows that the use of a relatively diffuse, but proper prior provides an estimated differential privacy of 3.2, which corresponds to maximal posterior odds of about 25. We interpret this result as meaning that if the influential observations that we actually deleted correctly depict those data rows that are most likely to change the LMM EBLUPs. Then, sampling from the posterior distribution of the random effects and releasing one vector draw (an estimated random effect for each county) from that sample has empirical $\epsilon-$differential privacy of 3.2.

8 Conclusion

The applications of two differentially private methods for releasing estimates from linear mixed-effect models allow some clear conclusions. The differentially private MLE is

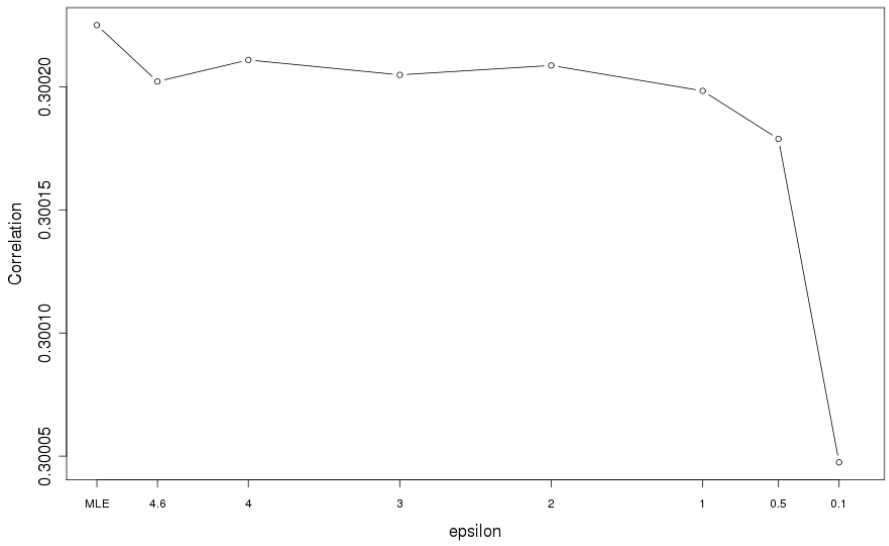

Figure 4: R-U Curve for JCR Linear Model with 100% ϵ budget for β

feasible in realistic problems when the random effects are limited to one high-dimensional factor, county in our case. For the protection levels that are feasible, the difference between the differentially private estimator and the MLE increases as the protection increases, as shown in our R-U plots. Our problem was chosen to give the differentially private MLE a reasonable chance of success. In particular, the dependent variable is bounded, which is not usually the case in detailed tabulations of continuous data–as routinely occur in small area estimation or detailed industry data. The differentially private MLE is not likely to work well for cases where there are several factors with many levels, as would be the case in our example if we used both county and detailed industry effects. We illustrated this failure with by adding seasonal county interactions. The differentially private MLE was not estimable for ϵ values less than three with seasonal county interactions, less than one for county and main industry effects, and less than 0.1 for main industry effects only.

The application of the Bayesian LMM to empirically estimate the differential privacy produced by a diffuse but proper prior gave very encouraging results. This method is a computational brute-force procedure that directly estimates an empirical analogue of ϵ. It is both feasible and practical for problems of the same degree of complexity as the ones in which the DP MLE was feasible, but the procedure may also be useful for more complex problems because the BLMM with a proper prior is not as delicate as the differentially private MLE computed using the sub-sampling method, which is limited by the number of sub-samples to models that are not as complex as the ones that can reasonably be fit with the BLMM.

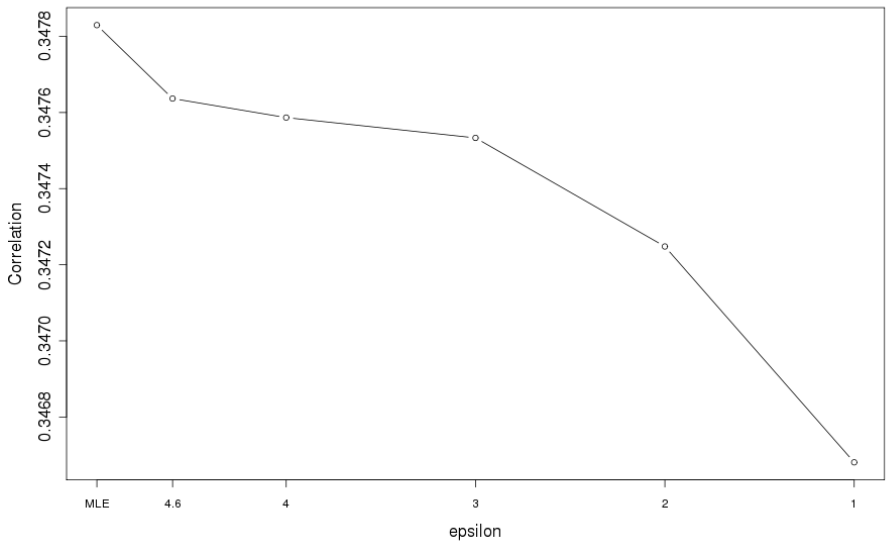

Figure 5: R-U Curve for JCR Linear Mixed Model with 88% ϵ budget for β and 10% for u

Acknowledgment

The authors acknowledge financial support from NSF grants BCS 0941226, SES 9978093, ITR 0427889, SES 0922005, and SES 1131848. They are also grateful for the comments of the JPC editor and referees.

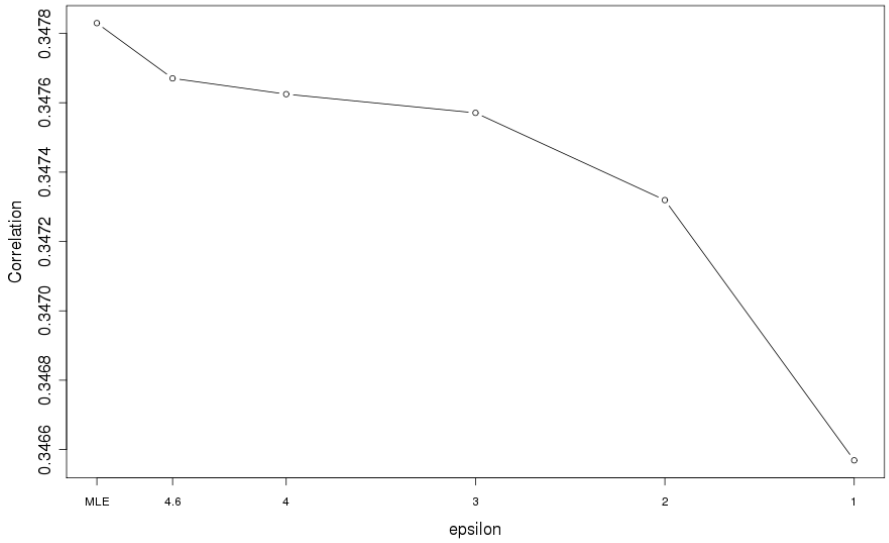

Figure 6: R-U Curve for JCR Linear Mixed Model with 10% ϵ budget for β and 88% for u

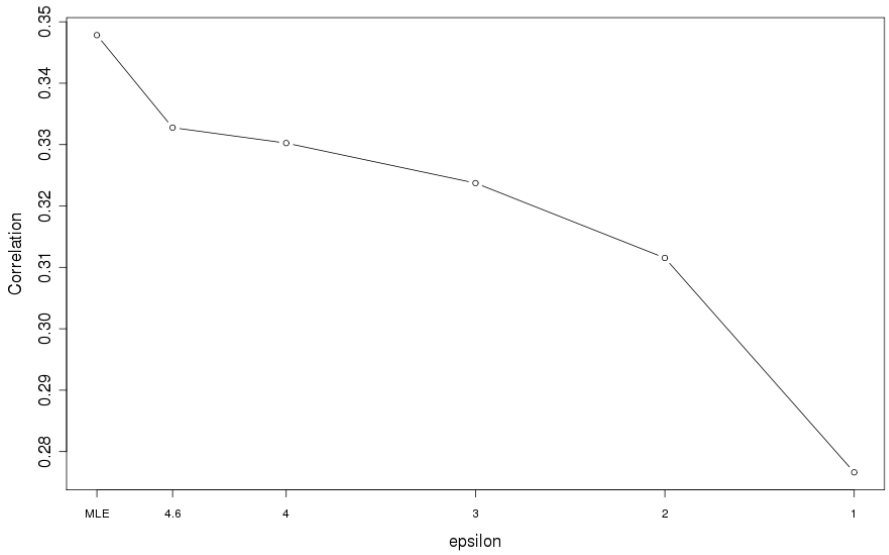

Figure 7: R-U Curve for JCR Linear Mixed Model with 97% ϵ budget for β and 1% for u

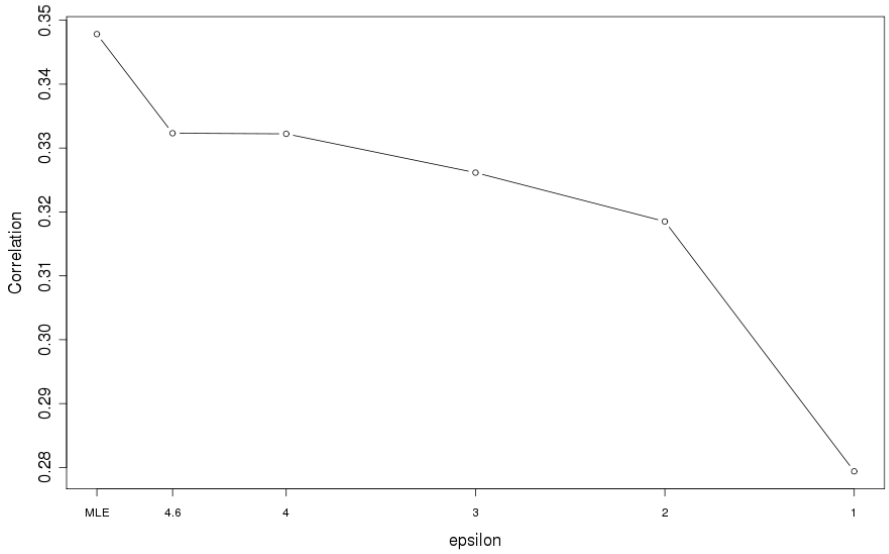

Figure 8: R-U Curve for JCR Linear Mixed Model with 1% ϵ budget for β and 97% for u

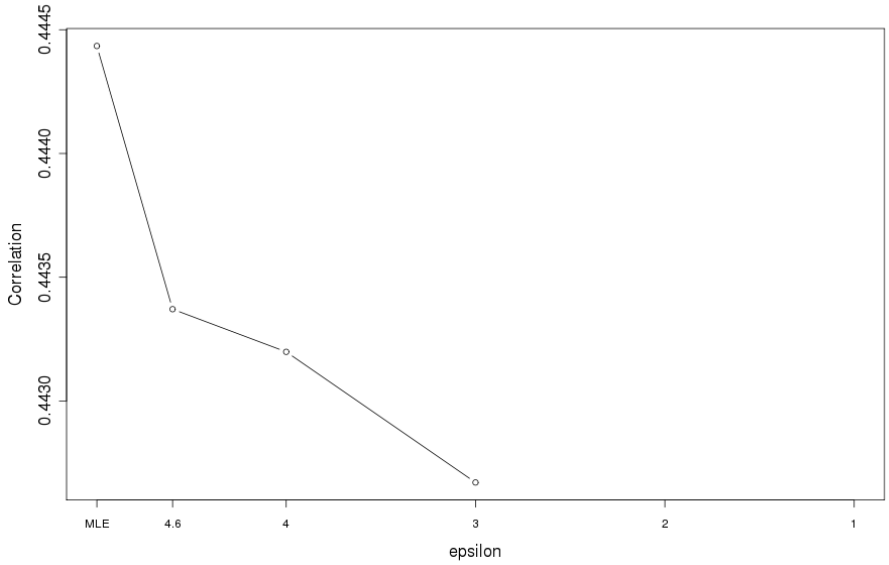

Figure 9: R-U Curve for JCR Linear Mixed Model with 88% ϵ budget for β and 5% for u and 5% for interactions

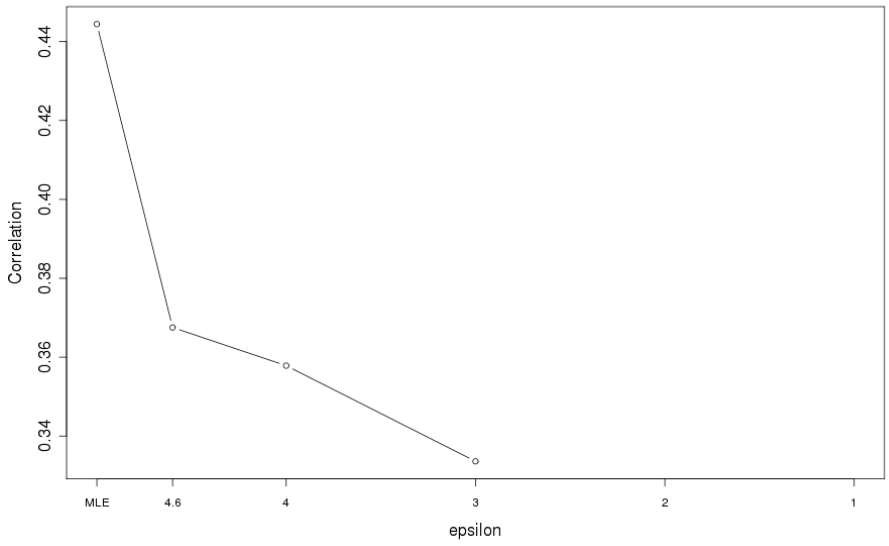

Figure 10: R-U Curve for JCR Linear Mixed Model with 97% ϵ budget for β and 0.5% for u and 0.5% for interactions

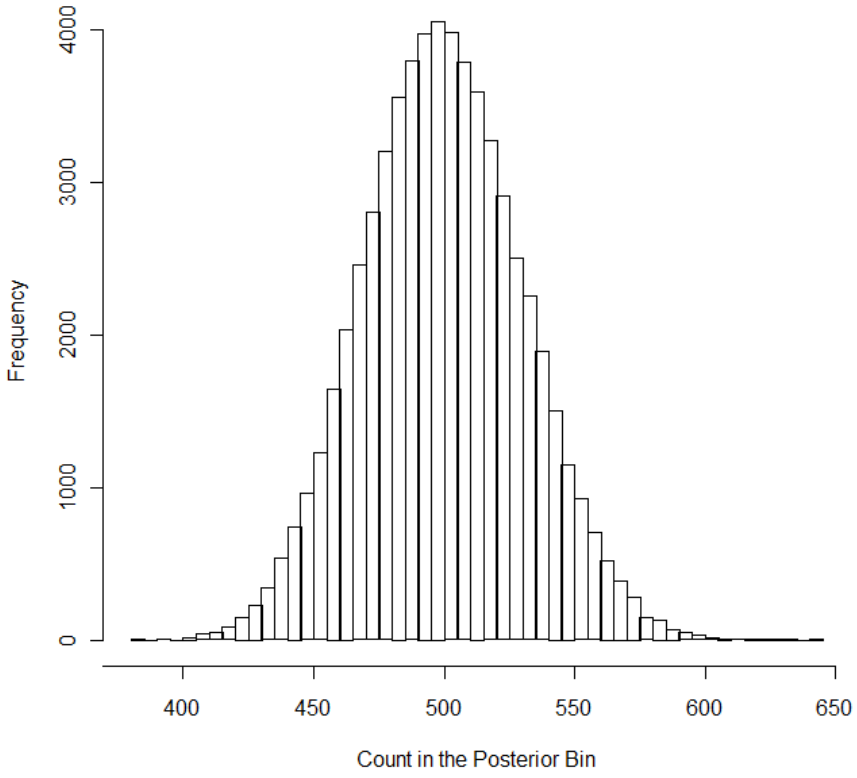

Figure 11: Histogram for the Replicated Model Including All Observations and Counties

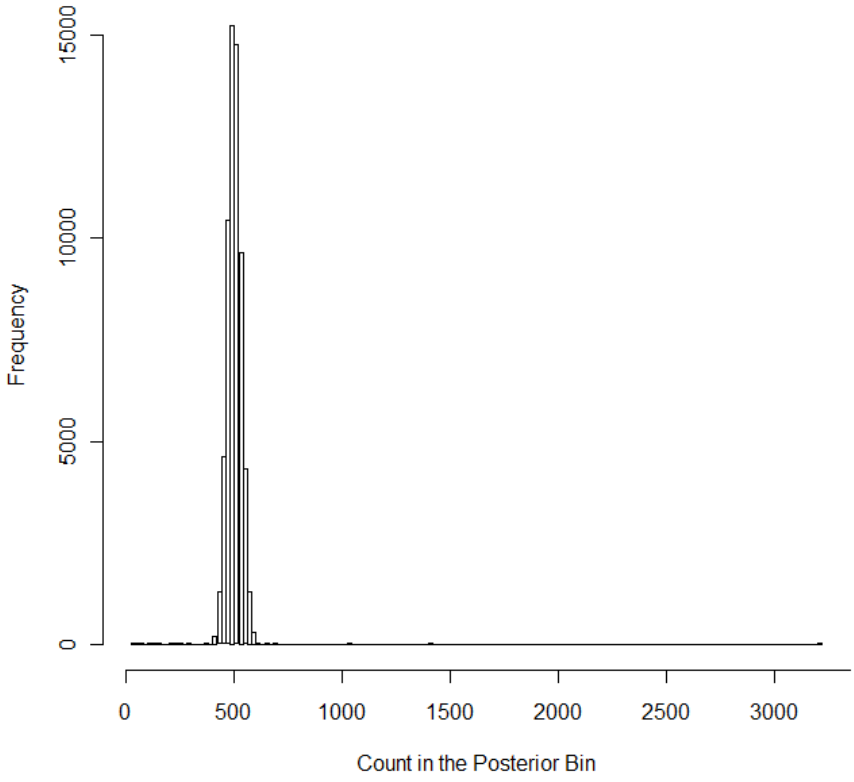

Figure 12: Histogram for the Model Deleting an Observation from County 3047

References

[1] Abowd, J., Stephens, B., Vilhuber, L., Andersson, F., McKinney, K., Roemer, M., and Woodcock, S. (2009). The LEHD infrastructure files and the creation of the quarterly workforce indicators. In T. Dunne, J. Bradford, and M. Roberts, eds., *Producer Dynamics: New Evidence from Micro Data*. Chicago, IL: University of Chicago Press for the NBER. 149–230.

[2] Abowd, J. and Vilhuber, L. (2011). National estimates of gross employment and job flows from the quarterly workforce indicators with demographic and industry detail. *Journal of Econometrics*, 161:82–99. doi:10.1016/j.jeconom.2010.09.008.

[3] Bates, D. (2004). Sparse matrix representations of linear mixed models. R Development Core Team.

[4] Bates, D. and Debroy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, 91(1):1–17.

[5] Bates, D. and Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-35.

[6] Chaudhuri, K., Monteleoni, C., and Sarwate, A. (2011). Differentially private empirical risk minimization. *Journal of Machine Learning Research,* 12:1069–1109.

[7] Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistical Review,* 15:429–444.

[8] Debroy, S. and Bates, D. (2003). Computational methods for single level linear mixed-effects models. Technical Report No. 1073, Department of Statistics, University of Wisconsin.

[9] Duncan, G., Elliot, M., and Salazar-Gonzalez, J. (2011). *Statistical Confidentiality: Principles and Practice*. New York, NY: Springer.

[10] Dwork, C. and Lei, J. (2009). Differential privacy and robust statistics. In *Proceedings of the 41^{th} ACM Symposium on Theory of Computing (STOC 2009)*, 371–380.

[11] Dwork, C. and Smith, A. (2009). Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2).

[12] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis Second Edition*, New York, NY: Chapman and Hall CRC.

[13] Hadfield, J. (2010). MCMC methods for multiresponse generalized linear mixed models: The MCMCglmm R packages. *Journal of Statistical Software*, 33(2):1–22. `http://www.jstatsoft.org/v33/i02/`.

[14] Hadfield, J. (2012). MCMCglmm course notes. Comprehensive R archive network. `http://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf`.

[15] Henderson, C., Kempthorne, O., Searle, S., and von Krosigk, C. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15:192–318.

[16] Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE 2008)*, 277–286.

[17] McCulloch, C. and Searle, S. (2001). *Generalized, Linear, and Mixed Models*, New York, NY: John Wiley & Sons, Inc.

[18] Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th AFM Symposium on Theory of Computing (STOC 2007)*, 75–84.

[19] Nobre, J. A. and Singer, J. D. M. (2007). Residual analysis for linear mixed models. *Biometrical Journal,* 49(6):863-875.

[20] Pinheiro, J. and Bates, D. (2000). *Mixed Effects Models in S and S-Plus*. New York, NY: Springer-Verlag.

[21] Prasad, N. and Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association,* 85(409):163–171.

[22] Rao, J. (2003). *Small Area Estimation*. Hoboken, NJ: John Wiley & Sons, Inc.

[23] Schabenberger, O. (2004). Mixed model influence diagnostics. In *Proceedings of the SAS Users Group International Conference (SUGI 29)*, Paper 189-29.

[24] Searle, S., Casella, G., and McCulloch, C. (1992). *Variance Components*. New York: John Wiley & Sons, Inc.

[25] Smith, A. (2008). Efficient, differentially private point estimators. Preprint arXiv:0809.4794v1.