THREE PAPERS ON TIME SERIES FORECASTING AND DATA PRIVACY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Matthew J. Schneider

May 2014

THREE PAPERS ON TIME SERIES FORECASTING AND DATA PRIVACY

Matthew J. Schneider, Ph.D.

Cornell University 2014

The first paper applies receiver operating characteristic (ROC) analysis to micro-level, monthly time series of the M3-Competition. Forecasts from competing methods were used in binary decision rules to forecast exceptionally large declines in demand. Using the partial area under the ROC curve (PAUC) criterion as a forecast accuracy measure and paired-comparison testing via bootstrapping, we find that complex univariate methods perform best for this purpose. The second paper develops a multivariate forecasting model designed for forecasting the largest changes across many time series. Using the partial area under the curve (PAUC) metric, our results show statistical significance, a 35 percent improvement over OLS, and at least a 20 percent improvement over competing methods. The third paper considers a particular maximum likelihood estimator (MLE) and a computationally intensive Bayesian method for differentially private estimation of the linear mixed-effects model (LMM) with normal random errors. The differentially private MLE performs well compared to the regular MLE, and deteriorates as the protection increases for a problem in which the small-area variation is at the county level. The direct Bayesian approach for the same model uses an informative, reasonably diffuse prior to compute the posterior predictive distribution for the random effects and the empirical differential privacy is estimated.

BIOGRAPHICAL SKETCH

Matthew J. Schneider is a Ph.D. Candidate in the Department of Statistical Science at Cornell University. He also holds a M.S. in Statistics from Cornell University, a M.S. in Public Policy and Management from Carnegie Mellon University, and a B.S. in Quantitative Economics from the United States Naval Academy. Previously, he was a Lieutenant in the U.S. Navy and policy analyst at the RAND Corporation.

ACKNOWLEDGEMENTS

I acknowledge NSF grants BCS 0941226, SES 9978093, ITR 0427889, and SES 0922005 for the funding of this work. I would also like to thank my coauthors who provided an equal contribution to each of these papers. They include Wilpen Gorr for the first two papers, and John Abowd and Lars Vilhuber for the third paper. Wilpen Gorr also served as my master's advisor at Carnegie Mellon University and Lars Vilhuber served as an informal advisor at Cornell. Additionally, I really appreciate the insightful comments and guidance from my committee at Cornell University over the years. My committee members include Martin Wells, James Booth, Sachin Gupta, and John Abowd. John Abowd, based in the Department of Economics at Cornell University, is also my main advisor and has been extraordinary in his role advising me in order to complete the Ph.D. in Statistics. Martin Wells and James Booth, based in the Departments of Statistical Science and Biological Statistics and Computational Biology at Cornell University, trained me in advanced statistical theory and methods and provided their great help to me since day one of the program. Sachin Gupta, based in the Marketing Department at Cornell University, has been instrumental in translating my statistical training to the marketing research community and continues to serve as a mentor to me. I am sincerely grateful.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

LARGE-CHANGE FORECAST ACCURACY: REANALYSIS OF M3-COMPETITION DATA USING RECEIVER OPERATING CHARACTERISTIC ANALYSIS

This paper applies receiver operating characteristic (ROC) analysis to micro-level, monthly time series of the M3-Competition. Forecasts from competing methods were used in binary decision rules to forecast exceptionally large declines in demand. Using the partial area under the ROC curve (PAUC) criterion as a forecast accuracy measure and paired-comparison testing via bootstrapping, we find that complex univariate methods (including Flores-Pearce2, Forecast Pro, Automat ANN, Theta, and Smart FCS) perform best for this purpose. The Kendall tau test of dependency for PAUC and a judgmental index of forecast method complexity provides further confirming evidence. We also found that decision-rule combination forecasts using three top methods generally perform better than the component methods, although not statistically so. The top methods for forecasting large declines match the top methods for conventional forecast accuracy in the M3 Competition's micro monthly time series. So evidence from the M3 competition suggests that practitioners use complex univariate forecast methods for operations-level forecasting, both for ordinary and large-change forecasts.

Key Words: Forecasting, ROC, M3-Competition, Exceptions Reporting, Large-Change Forecast Accuracy [1] [2]

[1] Gorr, W. L. and Schneider, M. J. (2013). Large-Change Forecast Accuracy: Reanalysis of M3-Competition Data Using Receiver Operating Characteristic Analysis. International Journal of Forecasting, Vol. 29, Issue 2.

[2] I acknowledge NSF grants BCS 0941226, SES 9978093, ITR 0427889, and SES 0922005.

1.1 Introduction

According to the management by exception (MBE) principle (Taylor, 1911), operations-level staff should make resource-allocation decisions for production of goods or services under ordinary conditions; however, under exceptional conditions staff should defer to higher-level management. This approach makes the best use of top managers' limited time, allowing them to deal with the difficult cases and the broader lines of strategies and policy making. In the case of product or service demand forecasting, one type of exception is a forecasted large change from current demand. If a forecasted change exceeds a predetermined threshold level, then the demand forecasting system issues an exception report, calling for diagnosis by staff members and possible actions by upper management.

Gorr (2009) introduced receiver operating characteristic (ROC) curves as an accuracy framework for time series forecasting in support of MBE. The ROC framework analyzes the tails of forecast error distributions for exceptional demand conditions; whereas, traditional forecast error measures (such as the MAPE and MSE) place the most weight on the centers of forecast error distributions and are best suited for ordinary demand conditions.

The "gold standard" for assessment of a forecast method in this paper is actual change in demand, available ex post. For example, as a policy, managers may wish to review the top few percent of actual decreases (or increases) as defined by a cutoff quantile point of the gold standard distribution. If a decision rule's threshold is crossed (*i.e.*, the rule "fires") and identifies an actual large change, the result is a "true positive," otherwise it is a "false positive." Other

outcomes are "true negative" where both forecasted and actual change are ordinary and "false negative" where the actual change was large but forecasted change was ordinary.

Gorr (2009) defined gold standard values as those values extreme in regard to the standardized time series of data, for example, the top five percent of standardized time series values. We can refer to this definition as "absolute" because it references the entire time series, whereas the current paper's definition is "relative" because it references only the last historical data point of a time series. The absolute definition is preferable when there are large costs in adjusting from a baseline or average level of production. Here an example is neighborhood crime level where a flare up above the baseline crime pattern comes to the attention of news reporters. Increased fear and lost confidence in police by the public are large societal costs in addition to losses by crime victims.

The relative definition for time-series gold standards, introduced in this paper, is preferable in circumstances where there are high costs in changing the production technologies from current levels coupled with a potential for avoiding future costs of holding excess inventory or not meeting customer demands. Examples are when additional machines need to be set up for increased demand or employees must be laid off for decreased demand. Because the focus is on large changes from current production levels, the decision horizon must be for the very short term of one or two steps ahead and the first step ahead is the more important. For example, managers might expect large changes in six months or a year and be able to track and adjust to changes incrementally, but a large change in the next time period requires swift and substantial changes in plans.

This paper applies ROC analysis to M3-Competition data and its univariate forecast methods. A key question is whether complex univariate forecast methods perform better than simple ones under ROC measures, similar to the case of Gorr (2009) who compared complex multivariate models to simple univariate methods for short-term forecasting and found complex methods to be more accurate. Most of the literature in the past 30 years supports using simple univariate methods for ordinary conditions (*e.g.*, the M-competitions). This paper provides additional evidence that complex forecast methods are significantly more accurate than simple methods for exceptions forecasting, and specifically for univariate methods.

We also investigate whether a combination forecast leads to increased accuracy for exceptions forecasting. For forecasting ordinary conditions, combinations are averages or weighted averages of individual forecasts. In contrast, combination forecasts for exceptions use "or" or "and" logical connectors for individual-forecast-method decision rules. For example, the best-performing combination forecast method in this paper requires that any decision rule for component forecast methods fire (with "or" connectors) for the combination decision rule to fire.

Also new in this paper is application of the partial area under the ROC curve (PAUC) as the forecast accuracy measure for exceptions forecasting. Included is a statistical test for differences in PAUC using paired comparisons and accounting for correlated data. The total area under the ROC curve has the interpretation of being the probability that a decision rule will signal a randomly-chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006)

Section 2 provides a brief literature review of forecast error measures and competitions. Section 3 covers the experimental design for reanalysis of M3 data and Section 4 provides results. Finally Section 5 concludes the paper.

1.2 Literature Review

In this section we review the M3-Competition and its analysis of forecast accuracy, especially in regard to micro monthly time series. We also review the literature on statistical tests available for comparing ROC curves.

1.2.1 M3-Competition data and forecast methods

Operations and marketing managers forecast individual products or product families in an attempt to meet demand. Hence we limit this study to the micro, monthly time series of the M3-competition which best match this decision setting. While both the M1 and M3 competitions have micro time series we use M3 data in this paper. The M3 competition has a wider range of univariate methods, especially more complex ones than M1. Furthermore, Koning *et al.*. (2005) provides judgmentally-derived complexity ranks for the M3 forecast methods made by three forecasting experts, which we relate to forecast accuracy. We averaged the complexity ranks across the experts and rescaled ties to yield the average ranks in Table 1.1. To learn more about the forecast methods in Table 1.1, see Table 1.2 in Makridakis & Hibon (2000, p 456).

Forecast Method	Expert 1	Expert 2	Expert 3	Average Rank
Naive2	1	1	1	1.0
Single	2	2	2	2.0
Holt	3	3	3	3.0
Robust-Trend	4	4	5	4.3
Winter	6	5	5	5.3
Dampen	5	6.5	7	6.2
PP Autocast	8	6.5	8	7.5
Theta SM	7	8	9	8.0
Comb SHD	10	9	5	8.0
Theta	9	10	14	11.0
BJ Automatic	11.5	11	11.5	11.3
Autobox1	11.5	13	11.5	12.0
Autobox3	18.5	13	11.5	14.3
Autobox2	18.5	13	11.5	14.3
ARARMA	13	15	17.5	15.2
Smart Fcs	15	18	17.5	16.8
Flores-Pearce2	15	18	17.5	16.8
Flores-Pearce1	15	18	17.5	16.8
Forecast Pro	17	18	17.5	17.5
Forecast X	21	18	17.5	18.8
RBF	20	22	21.5	21.2
AutomatANN	22	21	21.5	21.5

Table 1.1: Average rank of experts' judgmental assessment of forecast method complexity

1.2.2 M3-Competition results

A major conclusion of the M3-Competition is "Statistically sophisticated or complex methods do not necessarily produce more accurate forecasts than simpler ones" (Hibon & Makridakis, 2000, p 458). Micro monthly data, however, are a case for which complex forecast methods are more accurate than simple ones. For example, Table 1.2 summarizes best and worst performing methods for micro monthly data according to four forecast error measures used in the M3-competition (Koning et al.. 2005). All of the best methods are complex except for Theta, which has mid-range complexity. All of the worst methods

are simple, except Box Jenkins methods which also have mid-range complexity. Apparently micro monthly time series data have patterns that complex methods are able to estimate and make good use of under ordinary conditions. The question is whether complex models are also better for exceptional conditions. For example, when change is in progress neural networks have pattern recognizers that can turn on model components selectively to capture and extrapolate the change and expert systems can switch to more reactive models.

Error Measure	Best Four Forecast Methods (in order)
sMAPE	SmartFCS, Theta, AutomatANN, ForecastPRO
Median sAPE	SmartFCS, Theta, AutomatANN, ForecastX
RMSE	Theta, SmartFCS, ForecastX, ForecastPRO
Error Measure	Worst Four Forecast Methods (in order)
sMAPE	Robust-Trend, Naive2, Single, ARARMA
Median sAPE	Robust-Trend, Naive2, ARARMA, Single
RMSE	Robust-Trend, Naive2, RBF, Autobox

Table 1.2: Best and worst forecast methods for M3 micro monthly time series data (taken from Koning et al., 2005)

1.2.3 ROC Statistical Tests

Cohen *et al.*. (2009) and Gorr (2009) provide reviews of ROC curves and analysis applied to time series data monitoring and forecasting respectively. Hence this section only summarizes the ROC literature in regard to additional material on statistical tests introduced in this paper for time series testing.

Area under curve (AUC) is the total area under an ROC curve over the entire false positive rate (FPR) range of 0 to 1. The higher the AUC, the better the forecasting method (or other test mechanism). AUC can be computed using the trapezoidal rule given a comprehensive set of FPR, TPR pairs or by computing

the nonparametric Wilcoxon statistic, as shown by Hanley and McNeil (1982). The Wilcoxon statistic can be used to calculate the standard error of the AUC for statistical tests (Hanley and McNeil, 1982). Alternatively, the standard error and AUC can be determined using the DeLong, DeLong, & Clarke-Pearson (1988) method.

Partial area under curve (PAUC) is the area under an ROC curve for a specified FPR range, generally starting at zero. In many situations, a decision maker has a maximum FPR threshold which he or she is not willing to exceed and PAUC represents this case. PAUC can be computed using the trapezoidal rule and bootstrapping can be used to compute its standard error.

Parametric and nonparametric statistical tests for comparing the AUCs of two ROC curves with correlated data are described by Hanley and McNeil (1983). Forecasting competitions generally have correlated data because alternative forecast methods are applied to the same cross section of time series. The available tests require standard errors calculated by the Dorfman and Alf (1969) maximum likelihood program or the Wilcoxon statistic and a correlation coefficient calculated by the Pearson product-moment correlation method or the Kendall tau rank correlation coefficient. Bootstrapping removes the need for a covariance estimate and accounts for correlated data (Janes, Longton, & Pepe, 2009).

1.3 Experimental Design

We used ROC analysis to study large-change forecast accuracy for one- and two-month-ahead forecasts. This section describes how we processed the 474

time series of the M3 micro monthly data to create empirical ROC curves. The time series tend to be declining at the forecast origin, so we focused on exceptional declines—certainly a major concern of managers in firms selling products or services.

Included in this section is how we standardized data to facilitate cross-sectional specification of decision rule limits as well as how we tabulated results to produce ROC curves.

1.3.1 Standardizing forecasted change and its gold standard

Following is notation for the time series, forecasts, time series changes, and forecasted change.

Cross section of actual time series:

$Y_{it}(i = 1, \ldots I; t = 1, \ldots T+m)$ where i is a time series, t is time; T is the single, fixed forecast origin of the M3-Competition; and m is the forecast horizon (here we use $m = 1$ and 2 only)

Set of alternative forecast methods $j = 1, \ldots, J$ and forecasts:

$F_{ijt}(i = 1, \ldots I; j = 1, \ldots J; t = T + m)$

Forecasted change:

$ForecastDelta_{ijT+m} = F_{ijT+m} - Y_{iT}(i = 1, \ldots I; j = 1, \ldots J; m = 1 \text{ or } 2)$

The gold standard for comparison with forecasted changes is the true, ex post value for a one-month-ahead or two-month-ahead forecast minus the last

realization in the estimation data set:

$$Delta_{iT+m} = Y_{iT+m} - Y_{iT} (i = 1, \ldots I; m = 1 \text{ or } 2)$$

We need to standardize each ForecastDelta and Delta to remove scale and control variation, analogous to computing z-scores. Then we can use the same standardized threshold values of decision rules for each time series (as is done with t-statistics or normal tables). While we can estimate the sample mean and standard deviation for Deltas, there is a limitation in standardizing Forecast-Delta. The M3-competition had a single forecast origin for each time series and a single set of corresponding forecasts for $m = 1, \ldots, 18$, so there is only a sample of size one for each F_{ijT+m}. If the competition had used a rolling or expanding horizon design with many forecast origins, we could estimate the mean and standard deviation of ForecastDelta for each series and m. However for the M3-Competition we must use an approximation, which is facilitated by the way ROC curves are constructed. We need only assume that ForecastDeltas are proportional to Deltas by forecast method in the sample of time series. Then we can normalize ForecastDeltas by the mean and standard deviation of the deltas (see the Appendix).

1.3.2 Gold standard cutoff

Deltas that were a specified number of standard deviations below the mean were considered true large change values ("positives" in regard to ROC). We specified three cutoffs of -1.28, -1.65, and -2.33 standard deviations below the mean corresponding to 10%, 5%, and 1% quantile points of the delta distribution if it were normally distributed. For first differences ($Delta_{iT+1}$) of the 474

time series, there are 110, 74, and 24 positives for the 10%, 5%, and 1% cutoffs respectively. For second differences ($Delta_{iT+2}$), the corresponding number of positives are 89, 57, and 27. Even after removing spurious, regression-to-the mean cases in Section 3.3, the numbers of positives are higher than for a normal distribution because of the "fat" lower tail (and thin upper tail) of the distribution and also because our standardization is approximate. Regardless, it is important to analyze more than one gold-standard cutoff to examine how forecast performance varies with the definition of positives. For example, Gorr (2009) found ROC performance to improve with more extreme definitions, likely because the most extreme cases are the easiest to distinguish from the rest of the distribution.

1.3.3 Regression to the mean

It is necessary to control for regression-to-the-mean behavior in the case when exceptional values do not persist (*i.e.*, they are outliers) and time series patterns return to the mean of the series. For large declines, the problem occurs when the time series has a high outlier that returns to the mean. Take the case of one-step-ahead forecasts. Any non-responsive forecast method or model has good performance for the data point following the outlier, spuriously inflating AUC or PAUC measures. The actual data point returns to the mean while the unresponsive forecast method never left the mean. So $ForecastDelta_{ijT+1}$ fires a decision rule, testing positive, and $Delta_{iT+1}$ is a positive yielding a spurious true positive. The same is true for m-step-ahead forecasts, m steps after an increasing outlier. The one-step-ahead forecasts for the three thresholds of -1.28, -1.65, and -2.33 have 47 out of 110, 17 out of 74, and 3 out of 24 regression cases

respectively. The two-step-ahead forecasts have 30 out of 89, 15 out of 57, and 3 out of 23 regression cases. We excluded forecasts, and therefore corresponding time series, affected by regression to the mean from our analyses.

1.3.4 Forecast performance

For every threshold, standardized ForecastDeltas less or equal to the z-value threshold were considered to signal a large decrease (test positive). Forecasts methods with test positives in a series that had an actual positive are true positives. Otherwise, the forecast method provided a false positive. This process was repeated for a maximum of 475 z-value thresholds occurring at the boundaries of the 474 ranked normalized ForecastDeltas, thus spanning all possibilities for the construction of ROC curves.

True Positive Rates (TPRs, number of true positives divided by number of positives) and False Positive Rates (FPRs, number of false positives divided by number of negatives) were computed to obtain increasing two-dimensional points (FPR, TPR) for each method. The connection of these points created each method's empirical ROC curve and statistical tests were applied from Section 2.

1.4 Results

We decided to limit analysis to the PAUC measure for false positive rates between 0.0 and 0.2, believing that this would include the range in which most managers would be comfortable operating. Note that it is common in practice to use larger false positive rates (which are the same as type I error rates) than

used in theory testing (*e.g.*, see Cohen, et al.., 2009) depending on the cost of false negatives, prevalence of positives, and resources available for diagnosis and follow-up to test positives.

1.4.1 Partial Area Under Curve

We compare large-change forecast performance of forecast methods using a non-parametric bootstrap approach for paired comparisons between PAUCs. One sided p-values were computed for each PAUC threshold's top performing method to see whether it was statistically better than other methods. We use 1,000 bootstrap samples for each pair of methods.

See Table 1.3 for results at the 0.05 significance level for one-step-ahead forecasts over the FPR range of 0.0 to 0.2. We only included methods in the comparison that have complexity scores in Table 1.1 (dropping AAM1 and AAM2) and in addition we dropped Rule-Based Forecasting because it was designed for annual data and we are analyzing monthly data.

In general, complex methods performed significantly better than simple methods. Automat ANN, Flores-Pearce2, Forecast Pro, Smart FCS, and Theta were in the set of methods either best or not significantly different from the best for all three cutoff points used for gold standards. All but Theta are complex methods with subjective scores from Table 1.1 at 16.8 or higher. Theta has mid-range complexity with a score of 11.0. SmartFcs, with a complexity score of 16.8, was in the significantly better methods for the 95% and 90% gold standard cutoffs, and BJ automatic (mid-range complexity score of 11.3) joins the significantly better set for the 90% gold standard cutoff. Note in Table 1.3 that the

more extreme the positive cases are (*i.e.*, the more stringent the gold standard cutoff), the better the PAUC performance which is similar to findings by Gorr (2009).

99% gold standard	PAUC	p-value	95% gold standard	PAUC	p-value	90% gold standard	PAUC	p-value
Flores/Pearce 2	0.145		Theta	0.107		Automat ANN	0.071	
Automat ANN	0.123	0.112	Automat ANN	0.104	0.396	SmartFcs	0.069	0.431
Theta	0.122	0.086	Forecast Pro	0.104	0.374	Theta	0.067	0.275
ForecastPro	0.125	0.082	SmartFcs	0.102	0.315	ForecastPro	0.065	0.225
Dampen	0.112	0.045	Flores/Pearce 2	0.100	0.210	Flores/Pearce 2	0.065	0.203
Single	0.104	0.024	Theta sm	0.088	0.017	BJ automatic	0.059	0.080
SmartFcs	0.119	0.021	BJ automatic	0.089	0.015	Forecast X	0.056	0.044
Autobox2	0.099	0.016	ARARMA	0.081	0.008	Naive 2	0.048	0.028
Holt	0.101	0.015	PP-autocast	0.089	0.005	ARARMA	0.053	0.026
PP-autocast	0.105	0.010	Forecast X	0.084	0.005	Dampen	0.054	0.021
Winter	0.099	0.010	Flores/Pearce 1	0.084	0.004	Flores/Pearce 1	0.054	0.018
ForecastX	0.100	0.008	Autobox2	0.078	0.002	PP-autocast	0.053	0.014
BJ automatic	0.098	0.003	Dampen	0.083	0.002	Autobox3	0.049	0.014
Flores/Pearce 1	0.099	0.003	Autobox3	0.075	0.001	Robust-Trend	0.043	0.012
ARARMA	0.085	0.003	Holt	0.072	0.000	Theta sm	0.052	0.012
Naive2	0.076	0.001	Winter	0.070	0.000	Autobox2	0.051	0.008
Theta-sm	0.089	0.000	Single	0.070	0.000	Holt	0.044	0.002
Autobox3	0.074	0.000	Autobox1	0.062	0.000	Winter	0.042	0.001
Autobox1	0.059	0.000	Naive2	0.053	0.000	Single	0.038	0.000
Robust -Trend	0.025	0.000	Robust-Trend	0.040	0.000	Autobox1	0.035	0.000

Table 1.3: Paired comparisons with the top forecasting method of PAUC for FPR range 0.0 to 0.2 using bootstrapping: one-step-ahead forecasts.

For the two-step-ahead forecasts, the significantly better forecasting methods (in decreasing order of PAUC values) are as follows:

99% gold standard: Automat ANN, Theta sm, Flores/Peacre2, ForecastPro, BJ automatic, Theta, SmartFcs, and Autobox2

95% gold standard: ForecastPro, Flores/Pearce 2, Theta sm, Theta, SmartFcs

90% gold standard: Theta sm, Flores/Pearce 2, ForecastPro, SmartFcs

Most of these methods were in the significantly better sets for one-step-

ahead forecasts but Theta sm and Autobox 2 show up as new for two-step-ahead forecasts. Flores/Pearce 2 and Forecast Pro are in every significantly better set while Smart Fcs is close behind in all but one of those sets.

A ROC curve is a plot of true positive rate (TPR) versus false positive rate (FPR) obtained by varying the threshold level of an exceptions decision rule. Figure 1.1 displays a selection of ROC curves for one-step-ahead forecasts and the 95% gold standard case from Table 1.3. ROC curves for other cases are similar qualitatively. Shown are three top-performing methods (all complex) and three simple smoothing methods. Also shown is the line representing a chance decision mechanism.

For a given FPR, the method with the highest ROC curve is best, having the highest TPR. At 0.01 FPR there is no difference in performance but by 0.05 FPR the complex methods have a TPR range of approximately 0.32 to 0.42 while the simple methods have a range of 0.15 to 0.28. At 0.10 FPR, the complex methods have a TPR range of 0.61 to 0.63 while the simple methods only have a range of 0.40 to 0.49. So the complex methods have much better performance than the simple methods. At the maximum FPR rate in Figure 1.1 the best method finds just over 80 percent of the positive cases (gold standard large decreases). Note that Dampen has better performance than Single or Holt because, as shown by Snyder & Koehler (2008), it "...possesses a special capacity to adapt to structural change without direct intervention."

1.4.2 Complexity

This section investigates the effect of forecast method complexity on ROC performance, measured by PAUC, over the M3 micro monthly time series. We eliminated Rule-Based Forecasting from the analysis because it is an annual time series method, whereas the micro-level data analyzed in this paper are monthly. We also dropped the Naïve method because it yields 0 change comparing forecasts to last historical value and the AAM1/AAM2 methods which were not ranked by the experts for complexity in Table 1.1. We expected the relationship between complexity and PAUC to be positive.

Table 1.4 contains the results of applying Kendall's tau with a two-sided test and 0.05 significance level in regard to the dependence of PAUC for the FPR range of 0 to 0.20 on average rank for complexity in Table 1.1. Cases included are the three gold-standard cutoffs for defining positives and one- and two-month ahead forecasts. Five out of six cases have significant tests at the 0.05 level or better, thus providing further evidence that the complex forecast methods are best for the large-change forecast accuracy for the M3 micro monthly time series

	99% gold standard	95% gold standard	90% gold standard
One step ahead	tau = 0.197 p-value=0.241	tau = 0.464 p-value =0.005	tau = 0.535 p-value =0.001
Two steps ahead	tau = 0.432 p-value =0.009	tau = 0.379 p-value =0.023	tau = 0.411 p-value =0.013

Table 1.4: Kendall tau test

1.4.3 Decision-rule combination forecasts

It is well-known that a simple average combination of methods' forecasts often forecasts more accurately than the component methods (*e.g.*, Clemen,

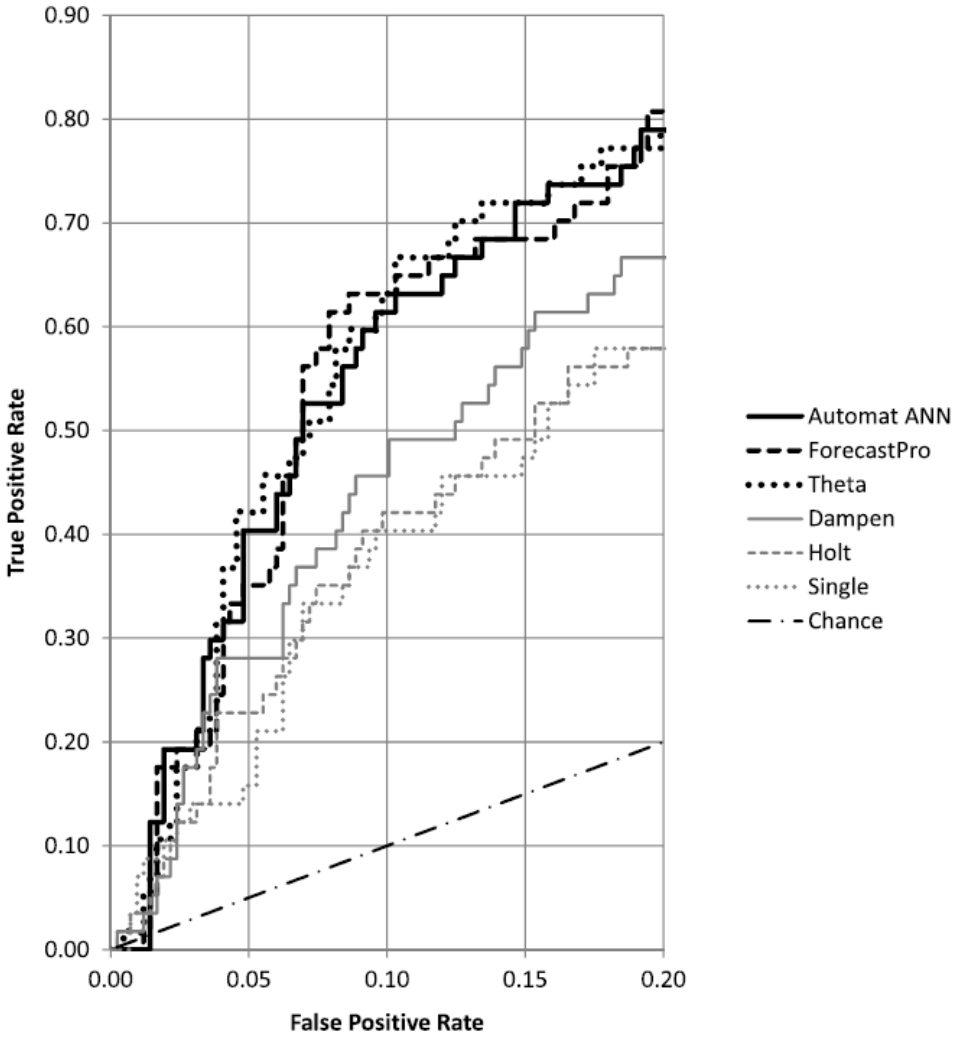

Figure 1.1: ROC curves for a selection of M3-Competition methods, micro monthly time series, one- month ahead forecasts, 95% gold standard cutoff.

1989). We propose combination forecasts for exceptions forecasting that combine decision rules instead of forecasts. For a decision-rule combination forecast with a fixed number of component forecast methods, if a prescribed number of component methods' rules fire (test positive), then the composite decision rule fires. The benefit of such a rule could be to make more conservative decisions, reducing false positives, or to be more inclusive and liberal adding more test

positives—depending on whether "and" or "or" logical connectors are used for component rules.

We created three combination forecasts, each with the same three top-performing, complex forecast methods: ForecastPro (expert system), Automat ANN (neural network), and Theta (decomposition method). Because each of the component methods have different modeling approaches, this combination promises to maximize information available for forecasting exceptions. The first combination rule (Min) has test positives whenever any of the three component methods has a test positive. The second (Median) is a median combination forecast that has test positives whenever two of the three component methods had a test positive. Finally, the third (Max) that has test positives when all of the top three methods have a test positive.

Table 1.5 has the results, using the same paired comparison test as in Table 1.3. Here we limit the comparisons to the three combinations as well as their three component forecast methods to see if combinations can improve forecast accuracy over components. The liberal Min combination is best for all three one-month-ahead cases and one of the three two-month cases, but is not significantly so at the 0.05 significance level. ForecastPro is best in the other two two-month-ahead cases. Thus rule-combination forecasts are promising: for forecasting large changes for important micro monthly time series, we recommend the Min rule-combination forecast.

One month ahead								
99% gold standard	PAUC	p-value	95% gold standard	PAUC	p-value	90% gold standard	PAUC	p-value
Min	0.134		Min	0.111		Min	0.075	
Median	0.128	0.302	Max	0.108	0.343	Max	0.070	0.188
ForecastPro	0.125	0.261	Median	0.108	0.328	Automat ANN	0.071	0.171
Theta	0.122	0.179	Theta	0.107	0.247	Theta	0.067	0.075
Max	0.120	0.135	ForecastPro	0.104	0.169	ForecastPro	0.065	0.040
Automat ANN	0.123	0.057	Automat ANN	0.104	0.079	Median	0.066	0.033
Two months ahead								
99% gold standard	PAUC	p-value	95% gold standard	PAUC	p-value	90% gold standard	PAUC	p-value
Min	0.141		ForecastPro	0.123		ForecastPro	0.092	
Median	0.136	0.343	Median	0.120	0.325	Min	0.090	0.389
ForecastPro	0.127	0.183	Min	0.119	0.347	Median	0.087	0.188
Theta	0.121	0.097	Theta	0.117	0.260	Theta	0.083	0.109
Automat ANN	0.129	0.078	Max	0.105	0.043	Automat ANN	0.077	0.076
Max	0.116	0.076	Automat ANN	0.094	0.013	Max	0.079	0.048

Table 1.5: Paired comparisons with the top forecasting method of PAUC for FPR range 0.0 to 0.2 using bootstrapping: one- and two-step-ahead forecasts for three rule-combination forecasts and their component forecast methods

1.5 Conclusion

This paper applied ROC analysis to the M3-Competition's micro monthly time series for one- and two-month-ahead forecasts. Using the partial-area-under-the-curve (PAUC) criterion, paired comparison testing via bootstrapping, and the Kendall tau we found that complex methods perform best for forecasting large declines in these time series, which tended to decline as a group over time. The classification of top methods matches that obtained using conventional forecast accuracy methods in the M3 Competition: complex methods forecast these series better than simple ones.

We also found that a rule-combination forecast, requiring that any of three decision rules of the combination methods fire to produce a test positive, to

perform better than the component methods but not with statistical significance.

Thus the evidence from the M3 competition suggests that operations managers should use complex methods such as Theta, a neural network, Forecast-Pro, or SmartFcs for forecasting both ordinary and large-change demand data.

CHAPTER 2

ROC-BASED MODEL ESTIMATION FOR FORECASTING LARGE CHANGES IN DEMAND

Forecasting for large changes in demand should benefit from different estimation than that used for estimating mean behavior. We develop a multivariate forecasting model designed for forecasting the largest changes across many time series. The model is fit based upon a penalty function that maximizes true positive rates along a relevant false positive rate range and can be used by managers wishing to take action on a small percentage of products likely to change the most in the next time period. We apply the model to a crime dataset and compare results to OLS as the basis for comparisons as well as models that are promising for large-change demand forecasting such as quantile regression, synthetic data from a Bayesian model, and a power loss model. Using the partial area under the curve (PAUC) metric, our results show statistical significance, a 35 percent improvement over OLS, and at least a 20 percent improvement over competing methods. We suggest managers with large numbers of time series (*e.g.*, for product demand) to use our method for forecasting large changes in conjunction with typical magnitude-based methods for forecasting expected demand.

Keywords: Management By Exception, PAUC Maximization, Large Changes, Forecasting Exceptions, ROC Curves. [1] [2]

[1] Coauthored with Wilpen L. Gorr

[2] I acknowledge NSF grants BCS 0941226, SES 9978093, ITR 0427889, and SES 0922005.

2.1 Introduction

Demand forecasting generally is done with extrapolative time series methods, such as exponential smoothing with level, trend, and seasonal components. Time periods during which the underlying univariate model is stable and forecast accuracy is acceptable are called "business as usual" (BAU) in this paper. Highly disaggregated time series, such as for product or service demand, however are notorious for having large changes—outliers, step jumps, turning points, etc.— that cannot be forecasted using simple extrapolative forecast models. Thus the time series forecasting field has long recognized the importance of handling exceptions to BAU; in particular, by developing time series monitoring methods for early detection of large changes (*e.g.*, Brown, 1959; Trigg 1964).

Time series monitoring supports reactive decision making, after large changes have occurred. Better, if forecast models are accurate enough, is to forecast large changes in demand to allow proactive decision making, with a chance of preventing losses or taking advantages of potential gains. This paper proposes a new estimation method for forecast models aimed at improving forecast accuracy for large changes in demand. The new estimation method minimizes a loss function based on a receiver operating characteristic (ROC) measure, partial area under the ROC curve (PAUC). Section 2 reviews the underlying decision framework—management by exception (MBE)—and ROC methods used to implement MBE for large-change demand forecasting, including PAUC. Essentially, ROC assesses predictions for binary classification, in the case of this paper that a future period will or will not have a large increase or decrease in demand.

While central-tendency forecast error measures, such as MAD, MAPE, and MSE, are best for evaluating forecast accuracy under BAU, recent research shows that a different kind of forecast error measure is needed for evaluation under large-change conditions for demand. Gorr (2009) compared univariate versus multivariate forecast models using the same data and found that forecast performance assessed using MAPE strongly favored simple univariate methods, whereas ROC assessment strongly favored the multivariate model. The multivariate model had leading indicator independent variables capable of forecasting large changes when there were large changes in lagged leading indicators. Gorr & Schneider (2013) compared simple versus complex univariate forecast models for large changes using monthly data from the M3 competition and found that PAUC assessments for large-change forecasts showed that the complex univariate models are significantly more accurate than simple univariate models. Apparently, complex univariate models have enough additional parameters in functional forms to make them sensitive to subtle indications of rapidly changing trends.

Parker (2011) goes a step further and shows that classification performance over seven measures of classification (which included AUC but not PAUC) is best by picking the right performance measures as a loss function for estimation. In line with this result, this paper provides evidence that parameter estimates for a multivariate forecast model made using the PAUC, ROC-based loss function are much more accurate for large-change demand forecasts than those from a central-tendency-based loss function (MSE). To our knowledge, there is no previous empirical research in time series forecasting using a ROC-based loss function for model estimation.

The general point of the emerging literature on large-change demand forecasting, including this paper, is that an organization can continue to use whatever extrapolative forecast models it prefers for BAU, but it needs a second, preemptive forecast model that takes over for large demand changes. Needed are two kinds of forecast accuracy measures (central tendency for BAU magnitudes and PAUC for classification), two kinds of forecast models (simple extrapolative for BAU and complex univariate or multivariate for MBE), and two kinds of loss functions for forecast model estimation (central tendency for BAU magnitudes and PAUC for MBE classification).

Section 2 provides motivation and background for the paper's new estimation method with an overview of MBE and ROC applied to large-change forecasting. Section 3 develops the ROC-based method for parameter estimation for a multivariate, leading indicator forecast model and develops comparison models. Section 4 describes the time series data used to calibrate the model and rolling horizon forecast experiment. Section 5 presents results comparing alternate forecast models with significance testing. Finally, section 6 concludes the paper with a summary and suggestions for future work.

2.2 Management by Exception for Demand Forecasting

This section provides an overview of MBE implemented with ROC analysis as applied demand forecasting. MBE and ROC provide a decision-making framework, model, and methods for managing exceptional, large-change demand conditions. The large literature on optimal inventory control models (*e.g.*, Brown, 1959) provides corresponding methods for BAU conditions, but mod-

els and methods for large-change conditions are fairly new. The key to success in this area is getting accurate forecasts, tuned for the tails of demand distributions, and that is the purpose of this paper, to the provide new estimation methods for large-change demand conditions.

MBE depends on the decision of whether or not to flag a forecast as being large-change, implemented via decision rules analogous to hypothesis testing, except that decision-rule thresholds cannot use the traditional Type I (false positive) error rates of empirical research (1 or 5 percent). Instead a false positive rate and corresponding decision-rule threshold must be determined based on cost/benefit considerations. ROC provides the decision model and methods for determining optimal false positive error rates, and corresponding decision rule thresholds.

MBE, one of the oldest forms of a management control system (Taylor, 1911; Ricketts & Nelson, 1987), provides the principle that only variances (exceptions) from usual conditions should be brought to managers' attention. All else should be handled by operational staff using standard procedures. Then managers' limited time can be devoted to decisions requiring their expertise and power on emerging problems or opportunities. One type of variance is a large change in demand for products or services (West et al., 1985; Gorr, 2009; Gorr & Schneider 2013). Demand is only partially affected by an organization's efforts, given competition in the market place, limits of marketing programs, and changing consumer tastes. Hence, large changes in demand are an important source of variances for triggering MBE reports to production and marketing managers. MBE can be reactive, based on detection with time series monitoring methods, or proactive, based on forecasting. First we discuss detection and then move on

to forecasting.

Time series monitoring methods, such as the Trigg (1964) and Brown methods (1959), compute a test statistic used in a decision rule analogous to hypothesis testing, for making the binary decision. The decision rule uses a threshold value, if exceeded by the test statistic, is a signal trip or "yes," otherwise the decision is "no" there is no large change. If "yes" then the time series undergoes diagnosis, using additional data and expertise, to determine if any intervention is needed into BAU practices (*e.g.*, a new marketing plan, price decrease, product improvements, or decreased production level). Next we discuss the mechanics of implementing such decision rules using ROC.

There needs to be an external determination as to when a time series has data points considered in fact to be large changes. Such a data point is called a "positive" and is determined by a "gold standard." All other time periods are "negatives." In public health, where ROC is used extensively (*e.g.*, Pepe, 2004), the analogous problem to MBE for demand forecasting is population screening for sick individuals. To be economically feasible, screening must use inexpensive and therefore imperfect tests, but then for individuals flagged as possibly sick, one needs a "gold standard" test for determining whether individuals are really sick or not, one that generally is expensive and more invasive. For example, for prostate and breast cancer screening, the gold standard is biopsy, examining sample tissue under a microscope. Biopsy is not infallible, but is much more accurate than screening tests such as PSA level in blood samples for screening prostate cancer.

Of course, the demand forecasting problem does not have gold standard tests, such as biopsy, for large demand changes. Instead, managers must use

judgment to determine changes large enough to be worth the cost of diagnosis and possible action. Gorr (2009) used a gold standard policy, that the top small percentage of large changes in standardized time series data be considered positives, and reasoned that police officials have means to make such judgments (*e.g.*, police would like to prevent the large changes that are reported in the news media). This gold standard is applied to out-of-sample forecasts during the evaluation stage, when actual values are available. A gold standard policy avoids the alternative of applying expertise and judgment to all time series points individually to determine positives in the evaluation phase of forecasting. Cohen *et al.* (2009) took this alternative, and while effective, was very costly.

There are four outcomes for a binary decision: true positive (the signal trips and the time period is a positive), false positive (the signal trips but the time period is a negative), false negative (the signal does not trip but the time period is a positive), and true negative (the signal does not trip and the time period is a negative). Application of a decision rule with a given threshold in repeated trials over time and across time series is summarized using a contingency (or confusion table) with frequency counts of all four possible outcomes. Common statistics from this table are the true positive rate, TPR = number of true positives/number of positives, and false positive rate, FPR = number of false positives/number of negatives. The complements of these statistics are the false negative rate and true negative rate.

It is a fact that increasing the true positive rate necessarily increases the false positive rate, so that there is a trade-off to be made in determining an optimal, corresponding decision-rule threshold. This is seen in the shape of the ROC curve, which plots true positive rate versus false positive rate for all possible

decision rule thresholds and is an increasing function with decreasing slope between (0,0) and (1,1). The higher the ROC curve for a model, the more accurate the binary decision model. An overall measure of the performance of a monitoring or forecast model thus is area under the ROC curve (AUC) which ranges between 0 and 1. Better in practice is partial area under the ROC curve (PAUC). This is the area for a restricted range of false positive rates, often from 0 up to 10 or 20 percent because the cost of processing signal trips for false positive rates exceeding those rates generally is excessive and/or beyond available resources. See Figure 2.1 in section 5 for example ROC curves and to get a sense of the kind of time series data being forecasted in this paper.

Empirical research uses traditional values, such as 1 or 5 percent, for false positive rates (Type I errors) that determine decision rule thresholds from normal or t-distribution tables. This practice implements a conservative view on accepting evidence of new theories. Business, however, needs to determine thresholds to obtain the optimal trade-off of true versus false positive rates. It is straightforward to write a utility model for the binary decision problem and to derive optimality conditions (*e.g.*, see Metz, 1978; Cohen et al., 2009). The optimal false positive rate is determined by finding the point at which a derived straight line is tangent to the ROC curve. The slope of that line depends on the prevalence of positives and the ratio of the utility of avoiding a false negative versus the utility of avoiding a false positive. For example, Pittsburgh police officials estimated that it is 10 times more important to avoid a false negative than a false positive when monitoring serious violent crimes for large increases, and this led to a 15 percent false positive rate as optimal for time series monitoring (Cohen et al., 2009). Likewise, population screening for prostate and breast cancers have false positive rates roughly in the range of 10 to 15 percent for most

parts of the world (*e.g.*, Banez et al., 2003; Elmore et al. 2002). In both crime and public health cases, the severe consequences of false negatives (not intervening when there is a large increase in serious violent crime or not catching cancer in early stages) outweighs the costs of processing false positives. So-called "A" items from ABC inventory analysis (*e.g.*, Ramanathan, 2006) are likely similar in terms of importance or consequence.

All of the framework and methods discussed in this section depend on being able to forecast large changes accurately enough. Thus the next section of this paper develops a model estimated using a loss function based on PAUC to best tune model parameters for MBE.

2.3 Multivariate Leading Indicator Modeling

Multivariate leading indicator models that restrict forecasting to linear predictors of the form

$$\hat{y} = X\hat{\beta} \tag{2.1}$$

are compared. y is the dependent vector with observations y_i for $i = 1, ..., n$ and X is the matrix of leading indicators (with time lagged values) with rows x_i. All models estimate $\hat{\beta}$ in-sample on different loss functions L which are functions of the data (y and X). Proposed models are well suited for large-change forecasting and central tendency models (ordinary least squares) provide a benchmark. First, the PAUC loss function is formally developed, then the proposed PAUC Maximization Forecast (PMF) model is developed, followed by

comparison models.

For all modeling, we define the initialization set as the set of data which is used to estimate $\hat{\beta}$ and not used in forecasting. The training set is is the set of data used for model selection based on pairwise comparison of out-of-sample results in the training set only. The test set is the set of data used to evaluate all models in this paper and report the results. This paper uses rolling horizon forecasts and iteratively conditions on all data up to time t to forecast data in time $t+1$. We define in-sample data as data up to time period t that is used to forecast out-of-sample data in time period $t+1$. Depending on the time period, in-sample data can exist in both the training and test sets, however pairwise comparisons for metaparameter selection are only performed in the training set (using the PAUC loss function as the comparison) and results are only reported for the test set. Both of these are done using out-of-sample forecasts only. See Table 2.2.

2.3.1 PAUC Loss Function

This section develops the functional form of the 1-PAUC loss function used for estimation. A manager states the gold standard policy that transforms the decision variable, $\mathbf{y}$, into a binary gold- standard vector, $\mathbf{g}$, where a 1 indicates a positive and a 0 indicates a negative,

$$\mathbf{y} \in \mathbb{R}^n \longrightarrow \mathbf{g} \in \{0,1\}^n \tag{2.2}$$

. A positive is an observation, worthy of investigation and possible inter-

vention, that we want to have flagged by a forecast. The policy is implemented using a threshold, not to be confused with decision-rule thresholds discussed below, for standardized values of the dependent variable, y^*, in our empirical application. Standardization matches the police criterion of equity for allocating resources to different regions of a city. If raw crime counts were used, all extra police resources would be allocated to the highest crime areas; whereas, with standardized crime counts any area, regardless of crime scale, with a relatively large increase in crime can get extra police resources. Section IV has details on the gold standard used in this paper.

ROC curves plot TPR versus FPR for all possible decision-rule thresholds of a given set of forecasts. ROC curves are constructed by comparing the rank of all forecasts to the gold standard vector. For forecast values $\hat{y}_i = X_i\hat{\beta}$, define the j^{th} decision rule threshold, $j = 1, 2, ...,$(1+ number of unique $\hat{y}_i$'s) corresponding to selected constants c_j's which divide the ranked $\hat{y}_i$'s. Then, a decision rule is defined under the j^{th} threshold and i^{th} observation where $\mathbf{1}_{\hat{y}_i > c_j}$ outputs a 1 if $\hat{y}_i > c_j$ or 0 otherwise where

$$DR_{i,j} = \mathbf{1}_{\hat{y}_i > c_j} \tag{2.3}$$

.

The resulting collection of TPRs and FPRs for all thresholds are

$$TPR_j(\hat{\beta}, X, \mathbf{g}) = (\sum_{i=1}^{n} \mathbf{1}_{DR_{i,j}=g_i=1})/(\sum_{i=1}^{n} g_i) \tag{2.4}$$

$$FPR_j(\hat{\beta}, X, \mathbf{g}) = (\sum_{i=1}^{n} \mathbf{1}_{DR_{i,j}-g_i=1})/(n - \sum_{i=1}^{n} g_i) \tag{2.5}$$

AUC is calculated as the sum of trapezoidal areas and PAUC is limited to a maximal FPR (*e.g.*, 20%) in practice.

$$AUC(\hat{\beta}, X, \mathbf{g}) = \frac{1}{2} \sum_{j=2}^{U} (FPR_j - FPR_{j-1})(TPR_j + TPR_{j-1}) \tag{2.6}$$

$$PAUC(\hat{\beta}, X, \mathbf{g}) = \frac{1}{2} \sum_{j=2}^{\{j:FPR_j \leq 0.20\}} (FPR_j - FPR_{j-1})(TPR_j + TPR_{j-1}) \tag{2.7}$$

1-PAUC is the loss function proposed in this paper for estimating forecast models used to implement MBE. Equivalent, of course, is PAUC maximization.

Explicit solutions for maximizing AUC exist under the assumption of normality (Su & Liu, 1993), but more recent research found that models which are tuned for AUC do not perform well for PAUC (Pepe et al., 2006; Ricamoto & Tortorella, 2010). PAUC maximization was recently studied in biostatistics for classifying patients as diseased or non-diseased using approximations to the PAUC function with wrapper algorithms (Wang & Chang, 2011) or boosting (Komori & Eguchi, 2010). Other biometric papers propose new PAUC maximization algorithms by using a weighted cost function with AUC and a normality assumption (Hseu & Hsueh, 2012). As such, the PAUC maximization papers concentrated on identifying the distributional differences between diseased and non-diseased populations, whereas, we use multiple time series as the dependent variable which presents challenges to boosting algorithms and sample size issues to PAUC approximations. Time series are treated as elements (versus

individuals as elements) and large changes within time series are positives (versus diseased individuals as positives). Our application differs structurally since large changes can occur in any time period with any time series.

2.3.2 PAUC Maximization Forecast Model

In this section, we detail the estimation procedure used to generate forecasts for our proposed PMF model. In overview, first, in each time period t, the proposed model chooses optimal coefficients β_t^* of the leading indicators X_t which have the best PAUC for the gold standard $\mathbf{g}_t$. Next, the estimation procedure combines current and past values of the optimal coefficients iteratively using an exponential smoothing procedure, which gives less weight to older estimates. This extra step provides consistency in parameter estimates from period to period. Finally, the proposed model forecasts large changes in time period $t+1$ and as time moves forward, the model is re-estimated for each successive set of forecasts.

For the current time period t, we define the cross-sectional loss as

$$L_t = 1 - PAUC_t(\mathbf{g}_t, X_t, \beta_t) \tag{2.8}$$

and select

$$\beta_t^* = \arg\min_{\beta_t} L_t \tag{2.9}$$

which minimizes L_t or equivalently maximizes $PAUC_t$ according to an

optimization procedure described below. $PAUC_t^*$ is calculated by using only functions of the in-sample vector $X_t\beta_t^*$. All unique cutoff values $c_1, c_2, ..., c_{(1+\text{number of unique values of } X_t\beta_t^*)}$ are chosen by first sorting across values within $X_t\beta_t^*$ and then averaging consecutive values which are not identical. These cutoff values represent various managerial decisions j of predicting large changes, $DR_{t,j} = \mathbf{1}_{X_t\beta_t^* > c_j}$. Then, $PAUC_t^*$ is estimated by inputting the vectors $DR_{t,1}, DR_{t,2}, ...DR_{t,(1+\text{number of unique values of } X_t\beta_t^*)}$ into the equations in the previous section.

To find optimal values of β_t^*, we employ the optim function in R, using the Nelder-Mead simplex method (which while is relatively slow, is known to be robust) for minimizing L_t (R Development Core Team, 2012). We set starting values equal to the OLS estimates of β_t and the run the optimization for a maximum of 500 iterations or until convergence. After L_t converges to a minimum, the current values of β_t are labeled β_t^* and the in-sample prediction vector $X_t\beta_t^*$ is determined to maximize $PAUC_t$.

Our early research using optim resulted in inconsistent parameter estimates from month to month. Thus, instead of using β_t^* for forecasting $t+1$, we train the forecasts over a rolling horizon of forecasts (*e.g.*, every month over several years). We incorporate a learning rate, λ, for the forecasting coefficients $\hat{\beta}_{t+1}$ which are a weighted combination of the current optimized values β_t^* and the past forecasting coefficients $\hat{\beta}_t$. Otherwise, our empirical results indicate that no past memory (*i.e.*, using only β_t^*, when $\lambda = 1$) results in and poor out-of-sample forecasts. We perform a grid search on the training set to determine the optimal $\lambda \in [0, 1]$ which represents the weighting of the optimization procedure in time period t.

$$\hat{\beta}_{t+1} = \lambda\beta_t^* + (1-\lambda)\hat{\beta}_t$$

The resulting forecast for time period $t+1$ and time series i with leading indicators $X_{i,t+1}$ uses only data from time period t or before and forecasts an index for a large-change:

$$\hat{g}_{i,t+1} = X_{i,t+1}\hat{\beta}_{t+1}$$

.

2.3.3 Comparison Models

The proposed PMF model is compared to several other models which incorporate the same multivariate leading indicators. The benchmark comparison model is Ordinary Least Squares (OLS) which we consider least suited to forecasting large changes. Other models appealing for large-change forecasting are also implemented. Power Loss models differ from squared error (*i.e.*, OLS) by varying the exponent of fit errors to give greater or less weight to extreme observations. Quantile regression fits the conditional quantiles (*e.g.*, median is 50%) of a given decision variable. Finally, a Bayesian technique is implemented using Markov Chain Monte Carlo (MCMC) techniques with the posterior predictive distribution (PPD) of the dependent variable. For notational simplicity, we drop the subscript t in this section. Although all model coefficients $\hat{\beta}$ are estimated on in-sample data, we choose the model metaparameters (p, τ, and quantile of the Bayesian regression) based on the PAUC loss function using out-of-sample data in the training set only. Further detail is given in the empirical application.

Power Loss

To estimate $\hat{\beta}$, we consider in-sample loss functions of the type

$$L = \sum_{i=1}^{n} |y_i - X_i\beta|^p$$

where $p \in [0, \infty]$. When $p = 2$, the solution solves the least squares problem, $\hat{\beta} = (X'X)^{-1}X'Y$, and the forecast $\hat{Y}$ is equal to the conditional mean, however, that interpretation is sacrificed here. Theoretically, as $p \to 0$, the loss is 0 when all $y_i = X_i\hat{\beta}$ (*i.e.*, perfect classification) and as $p \to \infty$, the loss is equal to the maximal observational loss over i. We select

$$\hat{\beta} = \arg\min_{\beta} L$$

for each p and use the results of a grid search on training data to determine the best p for out-of-sample forecasting. We expect that large values of p should perform well in-sample if there was only one large change since the prediction will minimize the maximal distance between y_i and $X_i\hat{\beta}$ over all i. Lower values of p give increasingly less weight to the maximal observational loss (*e.g.*, $p = 0.5$ penalizes each forecast by the square root of its distance to y_i).

Although we consider many power loss models for each p, we select the best power loss model with p^* according to the PAUC loss function on the training set. The resulting model with p^* is then evaluated on the test set.

Quantile Regression

Quantile regression estimates $\hat{\beta}$ by minimizing

$$L = (\tau - 1) \sum_{\{I: y_i < X_i\beta\}} (X_i\beta - y_i) + (\tau) \sum_{\{I: y_i \geq X_i\beta\}} (y_i - X_i\beta)$$

where $\tau \in [0, 1]$ and represents the τ^{th} quantile. We select

$$\hat{\beta} = \arg\min_{\beta} L$$

for each τ over an equally spaced grid of 101 values. When $\tau = 0.5$, the forecast $\hat{\mathbf{y}}$ is equal to the conditional median and powers loss when $p = 1$. Low and high values of τ represent extreme quantiles of conditional distribution of $\mathbf{y}$. Although there are a variety of quantile regression models for each τ, we select the best quantile regression model with τ^* according to the PAUC loss function on the training set. The resulting model with τ^* is then evaluated on the test set.

Quantile regression can also be interpreted as varying the ratio of costs of over-forecasting and under-forecasting. Quantile regression implicitly penalizes the costs of over-forecasting (when $y_i < X_i\hat{\beta}$) and under-forecasting (when $y_i \geq X_i\hat{\beta}$) by different ratios. This can be seen by setting $\tau = \frac{c_u}{c_u + c_o}$ where c_u is the cost of under-forecasting and c_o is the cost of over-forecasting. When c_u is small compared to c_o, τ represents a low quantile and $X_i\hat{\beta}$ will be small because an over-forecast is greatly penalized. In the case of forecasting large changes, it is not clear whether the cost of over-forecasting or under-forecasting should be of more importance since the performance of PAUC depends on the magnitude

and relative rank of the forecasts. In our empirical study, we seek to determine whether quantiles aligned with higher costs of over-forecasting perform better for PAUC because incorrect over-forecasts increase the false positive rate and therefore, decrease PAUC.

Bayesian Regression

One advantage of Bayesian estimation is that we can generate thousands of different forecasts for a single observation y_i and subsequently, analyze the distribution of these generated forecasts (synthetic data). From this distribution, we can select a quantile of the generated forecasts to forecast a large change. In the results section, we investigate whether forecasts based on quantiles perform better for MBE. Synthetic data models capture the underlying fit of the data and allow us to generate replicates of "fake data" using MCMC samples of the regression coefficients and error variance. So, it is possible to create thousands of artificial values for each y_i. The resulting synthetic data mimics the same distribution as y_i (*i.e.,* to include variation) because the data is generated conditional on $y_i|X_i, \beta, \sigma^2$ in the Bayesian model.

For the Bayesian regression, we use the same regression equation as OLS, $y_i = X_i\beta + \epsilon_i$, but place a diffuse but proper multivariate normal prior on β with mean zero and a block diagonal covariance matrix. We assume ϵ_i is independent and identically distributed for each observation and drawn from a normal distribution with mean zero and constant variance σ^2. For the prior of σ^2, we assume an Inverse-Wishart prior with an mean of zero and a degree of belief parameter of 1. We use MCMC techniques to sample draws of β and σ^2 from their resulting posterior distributions.

To generate the synthetic data, we use 1,000 posterior samples for each parameter (β, σ^2) after a burn-in of 1,000 samples. Since β is a k-dimensional vector, 1,000 samples are generated for each component which generates a k by 1,000 matrix. Values of y_i are generated 1,000 times for each i using the available samples. Then, those values are rank ordered and the appropriate quantiles are selected. The result is that the forecasted quantiles are taken on synthetic data generated from the conditional distribution of y_i (given X_i and parameter samples). Finally, we use a grid search of the empirical quantiles to select the optimal quantile for out-of-sample forecasting. The intuition is that forecasted quantiles other than the posterior mean or median may perform better for forecasting exceptional behavior.

Although there are a variety of Bayesian regression models for each quantile, we select the best quantile according to the PAUC loss function on the training set. The resulting model is then evaluated on the test set.

2.4 Empirical Application

2.4.1 Data Source

The data used in this paper are monthly crime counts by census tract from Pittsburgh, Pennsylvania. The dependent variable is the count of serious violent crimes (homicide, rape, robbery, and aggravated assault) while the 12 leading indicators are one, two, three, and four month time lags of illicit drug 911 calls for service, shots-fired 911 calls for service, and offense reports of simple assaults (Cohen et al., 2007; Gorr, 2009). The data span January 1990 through

December 2001 across 175 census tracts with 24,500 observations available out of 25,200 after dropping the beginning four month's observations used for time-lagged variables. For notation, we define y as the vector of violent crimes and X as the 12 column matrix of leading indicators. Table 2.1 shows the summary statistics for our data. Tract 404 represents a randomly selected low-crime area while tract 1115 is a random high-crime area. Besides overall performance of the computational experiment, we report forecast performance and gold-standard points for these two arbitrarily-chosen areas in the results section. Note that all crime counts are relatively low for monthly crime time series by census tract in Pittsburgh, making it challenging to obtain high forecast accuracy of any kind.

	Violent Crimes	Drugs	Shots	Assaults	Tract 404	Tract 1115
Min	0	0	0	0	0	2
Median	0	0	1	3	0	7
Mean	1.2	1.9	1.6	3.8	0.6	7.7
Std. Dev.	1.9	4.3	3.3	4.2	0.9	3.5
Max	29	71	50	42	4	20

Table 2.1: Summary Statistics of Crime Data

2.4.2 Gold Standard Policy

We employed a standardization procedure to define gold standard large changes in violent crimes (chosen to be about three percent of all census tracts) in accordance with Gorr (2009). In each census tract, the number of violent crimes were standardized according to their past smoothed mean and variance to account for the sizable time trends in the multiple-year data. The top five standardized values across all census tracts were labeled large changes each month as a gold standard policy defining positives.

In more detail, we perform a standardization procedure on each time series (*i.e.*, census tract) which shifts and rescales the current actual value in time t, y_t by its smoothed mean m_t and variance v_t, respectively. A low smoothing constant was used to allow the estimated mean to drift with the time series, but not to change appreciably from month to month. Smoothed means tend to yield data not over dispersed so that the Poisson assumption is valid. Thus we initialize values and assume $m_t = v_t$ from a Poisson distribution assumption since the number of violent crimes follows a count distribution. For each time period, we set our standardized value $y_t^* = \frac{y_t - m_t}{\sqrt{v_t}}$ and updated the estimates of the smoothed mean and variance by the current actual value. Once all values in each time series are standardized, we select the five largest values (three percent) for each month's cross-section of census tracts to define large increases in crime.

2.4.3 Rolling Horizons

Crime forecasting for deployment of police resources needs only one-step-ahead forecasts (one-month- ahead in this case). Urban police resources are highly mobile and easily and commonly reassigned or targeted. Also, most modern urban police departments have monthly review and planning meetings by sub- region (zone or precinct) so that one-month-ahead forecasts are needed. While forecasting large decreases in crime is perhaps useful for pulling police resources away from areas, the primary interest is crime prevention and forecasting large increases. A separate study, of the same magnitude and effort for large decreases, would be necessary for large decreases but is not conducted in this paper. A growing empirical literature shows that crime prevention in this

setting has at least moderate success (*e.g.*, Braga et al. 2012). We reestimate our models every month after forecasts are produced. All data up to time period t is used to forecast time period $t + 1$. Table 2.2 describes the conceptual setup.

Data Set	Initialization Set	Training Set		Test Set	
Data Used	In-Sample	In-Sample	Out-of-Sample	In-Sample	Out-of-Sample
Type	Parameter	Parameter	Metaparameter	Parameter	Not Applicable
PMF Model	β	β	λ	β	
Power Loss	β	β	p	β	
Quantile Regression	β	β	τ	β	
Bayesian Regression	β	β	Quantile	β	
OLS	β	β	Not Applicable	β	

Table 2.2: Parameters Optimized Across Data Sets

2.4.4 Forecast Evaluation

The PMF Model forecasts an ordinal index, $\hat{\mathbf{g}}$, where larger index values indicate that census tracts that are more likely to have large changes next month. Such a scale-invariant index is sufficient for use in decision rules for signalling large-change forecasts. On the other hand, all competing methods' have magnitude forecasts estimating demand and therefore forecasts need to be standardized according to their past mean and variance. Since standardization is not scale-invariant, this transformation changes each method's PAUC. If no standardization of magnitude-based methods were performed, competing methods would always forecast large changes for the most violent census tracts because their magnitudes are higher. Empirically, standardization improved PAUC performance for magnitude-based methods.

Our data consisted of 175 time series with 136 months each. We used 44

months of data to initialize each method in the initialization set, the next 24 months for the training set, the next 12 months to burn in the exponential smoothing procedure for the gold standard policy, and the final 60 months for evaluation in the test set. All model parameters were chosen via grid search in the 24 months of the training set. Grid searches were performed over 101 values of the learning rate λ for the PAUC Maximization Forecast model, the power p for the power loss model, the quantile τ for quantile regression, and the quantile of the PPD for the Bayesian model. Rolling horizon forecasting on out-of-sample data was performed in the training set to select the best values of these metaparameters. The last 60 months of data in the test set was used for evaluation of out-of-sample forecasting and the results are presented in the next section. Forecasts were made by each model with a rolling horizon of one-month, consistent with decision making in crime forecasting. All models were re-estimated at every forecast origin, however, only forecasts of the magnitude-based methods were standardized and their coefficients were not adjusted.

2.5 Results

We summarize all 60 months of out-of-sample forecasts for each model with a single ROC curve. Each ROC curve represents 10,500 forecasts (60 months times 175 census tracts) to forecast 300 large changes. The 300 large changes consisted of the top 5 large changes for each month. Smoothed ROC curves up to a false positive rate of 20% (PAUC's relevant range) are shown in Figure 2.1 where the PMF model is seen to strongly dominates competing methods. This means that the proposed model forecasts the most gold standard large changes at any given false positive rate. For example, this figure shows that the proposed

model has double the number of true positives at a false positive rate of 10%.

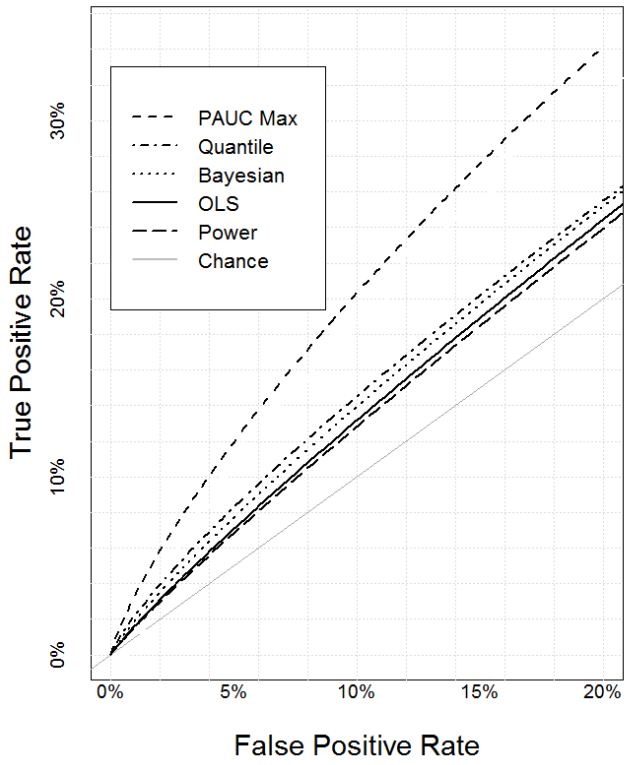

Figure 2.1: Smoothed ROC Curves Over Five Years in Test Set

Table 2.3 presents corresponding results for each model. In order to test for statistical significance between two correlated ROC Curves, we used 1,000 bootstrap samples with the R package pROC (Robin et al., 2012). We compared the PAUC of the proposed model to other models. Each bootstrap sample randomly selects the same number of large changes and observations as the original data, and 1,000 PAUCs are generated. Differences are calculated and their standard deviation divides the PAUC difference of two original ROC Curves to generate the test statistic found in the table. The PAUC Maximization Forecast model had a statistically greater PAUC than all other models at the 5% alpha level in the test set.

Model	PAUC	Test Stat.	One Sided p-value	Percent Improvement
PAUC Maximization	.0359	-	-	
Quantile Regression	.0297	1.74	.041	20.5%
Power Loss	.0285	2.05	.020	25.7%
Bayesian Method	.0283	2.04	.021	26.8%
OLS with Leading Indicators	.0266	2.51	.006	35.0%

Table 2.3: Five-Year ROC Curve for the Test Set

Using PAUC methodology, quantile regression models with low quantiles forecasted large changes more accurately than models with high quantiles in the training set. Specifically, we found that for forecasting the top three percent of large changes, quantile regression did the best in the training set when it implicitly assigned a cost of over-forecasting 12 times greater than a cost of under-forecasting. This represented the $\tau^* = .08$. Additionally, a Bayesian method based on empirical quantiles of sampled synthetic data and the power loss method had a higher PAUC than OLS, but not statistically so. Power Loss had poorer performance than OLS at higher powers in the training set. $p^* = .68$ was selected for use in the test set since it had the maximum PAUC in the training set.

We also evaluated forecasts with the continuous measure of correlation to the gold standard vector (*i.e.*, 1 if large change, 0 otherwise) in the test set. Table 2.4 shows that all correlations were significantly different than zero at the 5% alpha level and the PMF Model had the highest correlation. Correlations are all positive indicating that higher forecasts are more closely associated with large changes. However, correlations are very low because 97% of the gold standard vector was zero (*i.e.*, not a large change) as defined in our gold standard policy.

Over the five-year test set, we show the actual number of violent crimes for

Model	Correlation	p-value
PAUC Maximization	5.51%	0.011
Quantile Regression	2.41%	0.009
Power Loss	2.50%	0.014
Bayesian Method	2.56%	0.011
OLS with Leading Indicators	2.48%	0.000

Table 2.4: Five-Year Correlation of Gold Standard to Forecasts, Test Set

the two sample census tracts previously summarized in Table 2.1, in Figure 2.2. The black point markers in Figure 2.2 are large-increases positives from our gold standard policy. For Census tract 404, the PMF Model's highest index forecast (0.102) occurred in month 46 at the only large increase and therefore, our proposed method results in zero false positives for a decision rule using a cutoff of 0.102. However, OLS's standardized forecast during month 46 (0.113) is the 11th highest forecast among the five years. Therefore, if a manager used a cutoff of .113 for OLS, it would result in one true positive and ten false positives. In sum, for the Census tract 404, the PMF model outperforms OLS in terms of forecasting the large increase. On the other hand, Census tract 1115 is a high-crime area and the two large changes identified occurred in months 19 and 37, and had 15 and 16 violent crimes, respectively. The PMF Model forecasted the two large increases as the 29th and 30th highest forecasts in Census tract 1115, respectively. OLS's standardized forecasts placed these two large increases as the 21st and 40th highest forecasts for Census Tract 1115. Therefore, although OLS had less false positives for forecasting the first large increase, the PMF Model had less false positives (28) for forecasting both large changes compared to OLS (38).

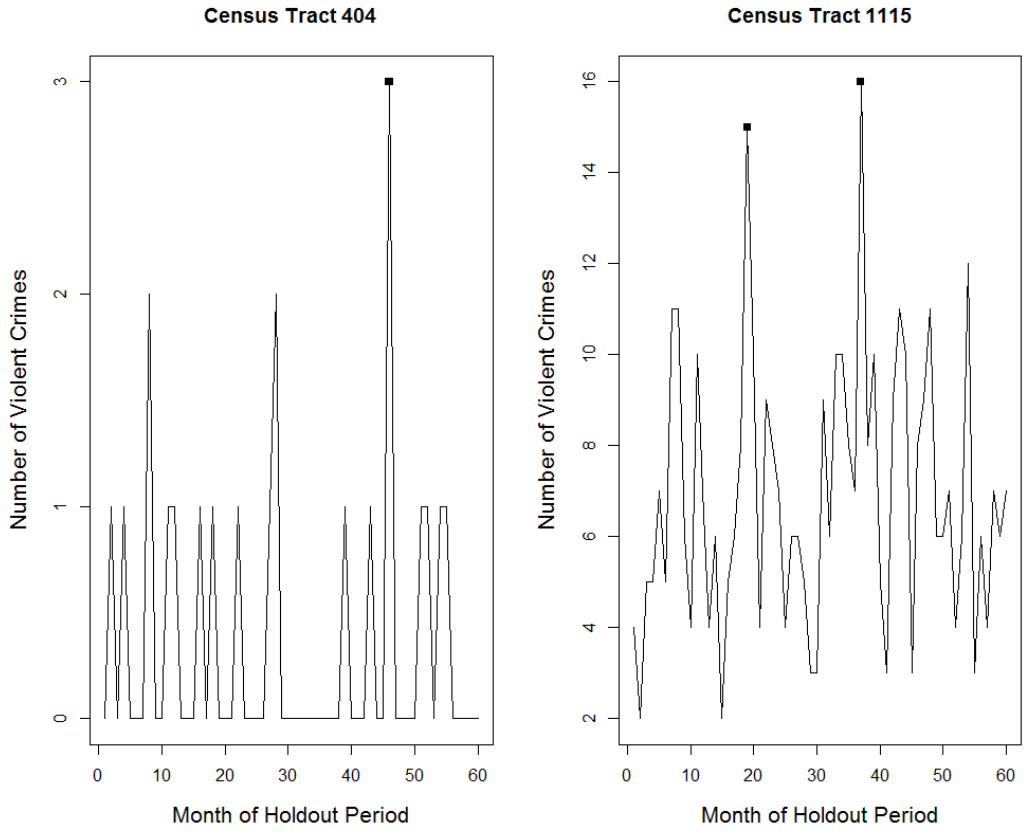

Figure 2.2: Actual Violent Crimes for Two Census Tracts in the Test Set

2.6 Conclusion

This is the first paper to use partial area under the curve (PAUC) from receiver operating characteristics (ROC) analysis as the basis for a loss function to estimate forecast model parameters (1 – PAUC is the loss function used). PAUC tunes forecast models to the tails of product or service demand distributions, thereby substantially increasing large-change forecast accuracy (with statistical significance) over models estimated using MSE or other central tendency measures as the loss function. The PAUC-based model is also statistically and substantially more accurate than other comparison models that can be tuned

or used for large-change forecasting. The forecast model of this paper is multivariate with leading indicators able to forecast large changes when there are large changes in the lagged indicators. The same loss function and optimization methods can be applied to any forecast model, including the complex univariate models found superior to simple univariate models for large change forecasting by Gorr & Schneider (2013). Our findings confirm previous research which says that models which include loss functions for the accuracy error metric desired perform the best (Parker, 2011).

Accurate large-change forecasting is the key to proactive management-by-exception (MBE) for managing product inventories and marketing programs. The MBE principle states that only variances (exceptions) should be brought to the attention of managers, with business-as-usual decisions handled by staff using standard procedures. Thus the decision to be made for MBE is binary, whether or not a forecast is large enough to bring to management for further analysis and possible interventions. Such decisions are made using decision rules analogous to hypothesis testing, however, it is necessary to select decision rule thresholds for MBE and ROC enables one to estimate and compute an optimal decision rule (and corresponding false positive rate).

The implications of MBE and demand forecasting is that firms should continue to use their current extrapolative forecast models for business-as-usual conditions, but also implement a large-change forecast model, such as developed in this paper. When a large-change decision rule fires for a demand time series, the business-as-usual forecast is preempted with management review of the product.

CHAPTER 3

DIFFERENTIAL PRIVACY APPLICATIONS TO BAYESIAN AND LINEAR MIXED MODEL ESTIMATION

We consider a particular maximum likelihood estimator (MLE) and a computationally intensive Bayesian method for differentially private estimation of the linear mixed-effects model (LMM) with normal random errors. The LMM is important because it is used in small-area estimation and detailed industry tabulations that present significant challenges for confidentiality protection of the underlying data. The differentially private MLE performs well compared to the regular MLE, and deteriorates as the protection increases for a problem in which the small-area variation is at the county level. More dimensions of random effects are needed to adequately represent the time dimension of the data, and for these cases the differentially private MLE cannot be computed. The direct Bayesian approach for the same model uses an informative, reasonably diffuse prior to compute the posterior predictive distribution for the random effects. The empirical differential privacy of this approach is estimated by direct computation of the relevant odds ratios after deleting influential observations according to various criteria.

Keywords: Differential Privacy; Statistical Disclosure Limitation; Privacy-Preserving Datamining; Linear Mixed Models; Quarterly Workforce Indicators; MLE; REML; EBLUP; Posterior Distribution; Random Effects. [1] [2]

[1] Abowd, J.M., Schneider, M.J., and Vilhuber, L. (2013). Differentially Private Applications to Bayesian and Linear Mixed Model Estimation. Journal of Privacy and Confidentiality, Vol. 5, Issue 1.

[2] I acknowledge NSF grants BCS 0941226, SES 9978093, ITR 0427889, and SES 0922005.

3.1 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 subsample and aggregate algorithm (Smith, 2008). The second approach produces differentially private linear predictors for regularized Empirical Risk Minimization (ERM) by perturbing an objective function (Chaudhuri et al., 2011). 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 ex-

pect 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 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 (Dwork & Smith, 2009) 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 datamining (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 governmental 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 (Duncan et al., 2011). 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 (Duncan et al., 2011; Dwork & Smith, 2009).

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 (Machanavajjhala, et al., 2008; Prasad & Rao, 1990; Rao, 2003). 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.

3.2 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 Abowd et al., 2009; Abowd & Vilhuber, 2011 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, S is separations within quarter. The identity $JCR - JDR = AR - SR$ holds for all entities and time periods (Abowd et al., 2009). 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 III 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.

3.3 Model Specifications

3.3.1 Linear Mixed Model

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 (McCulloch & Searle, 2001). 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 3.1 (Searle et al., 1992).

Linear Mixed Model Specification

Our statistical model is specified as follows:

$$y = X\beta + Zu + \xi \tag{3.1}$$

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 unique county within 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 (3.1) .

Henderson *et al.* (1959) 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.

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 (Bates & Maechler, 2010) and takes advantage of sparse matrix computational methods (Bates, 2004). Table 3.2 shows a summary of the REML estimates produced for our model. Initial global estimates are calculated from Table 3.2 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 IV and V 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 $\left(\beta, G/\sigma_\xi^2, \sigma_\xi^2\right)$ (Bates & Debroy, 2004; Debroy & Bates, 2003), Bates and Debroy (2004) 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 $\left(\beta, \sigma_\xi^2\right)$.

3.3.2 Bayesian Linear Mixed Model

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 empir-

ical 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 V.

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 (Gelman et al., 2004). 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.

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

$p(u_c|y, X, Z, \beta, \sigma^2_\xi, \sigma^2_c)$ for every county c. Section V has further details.

3.4 Differentially Private Estimation via Sub-sampling

We use LMMs and Smith's (Smith, 2008) 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 VI. We apply Smith's method of differential privacy via sub-sampling (Nissim et al., 2007; Smith, 2008) directly to the full data matrix from the QWI.

3.4.1 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 3.2 using the data for each block only.

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

McCulloch and Searle (2001) 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 (1990) 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 increased. 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

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

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

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

end

forall the *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 the *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

3.4.3 Averaging Sub-samples as the Aggregation Function

Average the estimates over k blocks:

$$\hat{\beta}^{**} = \frac{\sum_{i=1}^{k} \hat{\beta}_{(i)}}{k}$$

$$\hat{u}^{**} = \frac{\sum_{i=1}^{k} \hat{u}_{(i)}}{k},$$

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

and

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

Next, draw R_β^ϵ, $R_u^\epsilon$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 (Smith, 2008) or the maximum ranges of the parameters β, μ, and σ, respectively, as shown in Table 3.3. 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 IV-D. 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 the 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.[3] Therefore, we must keep k at a reasonable level and in Section V, 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 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 3.3. For JCR and JDR, all three components of $\hat{\sigma}$ are bounded since standard deviations should be a maximum of 0.5 for rates

[3]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."

$\in (0, 2)$. 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 3.3. 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 3.3 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 (Abowd et al., 2009), 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.

3.4.4 Number of Sub-samples

Smith (2008) 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 3.1 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 3.1.

3.4.5 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 (Smith, 2008).[4]

[4] $t\Lambda_{\beta_{21}} \leq 101(.01) = 1.01$ 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.

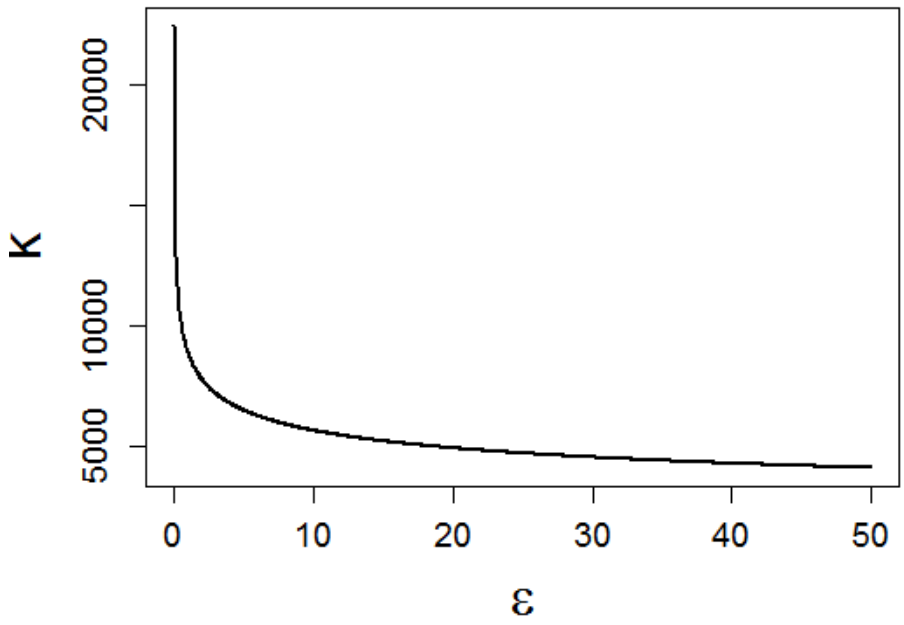

Figure 3.1: Equivalence table for optimal k over values of ϵ for JCR

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 (2008), this results in the Laplace random variables

$\hat{\beta}^{DP_\epsilon} = R_\beta^\epsilon + \hat{\beta}^{**}$ and $\hat{u}^{DP_\epsilon} = R_u^\epsilon + \hat{u}^{**}$ each being ϵ-differentially private where the Laplace noises were defined in Section IV-C. Define an arbitrary fitted value of the vector $\hat{y}^{DPC\epsilon} = X\hat{\beta}^{DP_\epsilon} + Z\hat{u}^{DP_\epsilon}$ as $\hat{y}_a^{DP_{c\epsilon}}$. $\hat{y}_a^{DP_{c\epsilon}}$ 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 3.3, 3.4, 3.5, 3.6, 3.7, and 3.8.

3.5 Differentially Private Estimation via Expected Risk Minimization

We use BLMMs for Chaudhuri, Monteleoni, and Sarwate's (2011) 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 approxi-

mation 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 (Chaudhuri et al., 2011). 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.

3.5.1 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 $\frac{\nu^2 V^2}{(\nu-2)^2\left(\frac{\nu}{2}-2\right)}$ 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 $\Sigma = 16^2 \frac{v^2 V^2}{(\nu-2)^2\left(\frac{\nu}{2}-2\right)} I$ 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 $\Sigma = 16^2\frac{v^2V^2}{(\nu-2)^2\left(\frac{\nu}{2}-2\right)}I$.

3.5.2 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 (Hadfield, 2010). 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 subsampling 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.

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," $H = X(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}$ (Schabenberger, 2004). 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 (Schabenberger, 2004). Nobre and Singer (2007) 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 (2000) 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 $\hat{u}$.

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 V-A 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

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

and

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

where the second equality is due to Bayes law. The expressions for M_1 and M_2 is 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 (1977). 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 $D\tilde{}i$.

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 $\max\left(\left|\ln M_1\right|,\ \left|\ln M_2\right|\right)$ by comparing the outcomes in the benchmark model, $P\left(u_c \in b|D\right)$ with outcomes in models estimated deleting an influential observation, $P\left(u_c \in b|D\tilde{}i\right)$, using a discretized posterior with 20 bins whose boundaries are determined by $P\left(u_c \in b|D\right)$.

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 $u_c|D\tilde{}i$ 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 $\epsilon = \max\left(\left|\ln M_1\right|,\ \left|\ln M_2\right|\right)$ where $M_1 = \max_{i,c,b}\left[\frac{n_{i,c,b}}{500}\right]$ and $M_2 = \min_{i,c,b}\left[\frac{n_{i,c,b}}{500}\right]$.

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 (Gelman et al., 2004). 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 3.2 below and found no issues with convergence.

3.6 Results

3.6.1 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 *et al.* (2011) states that

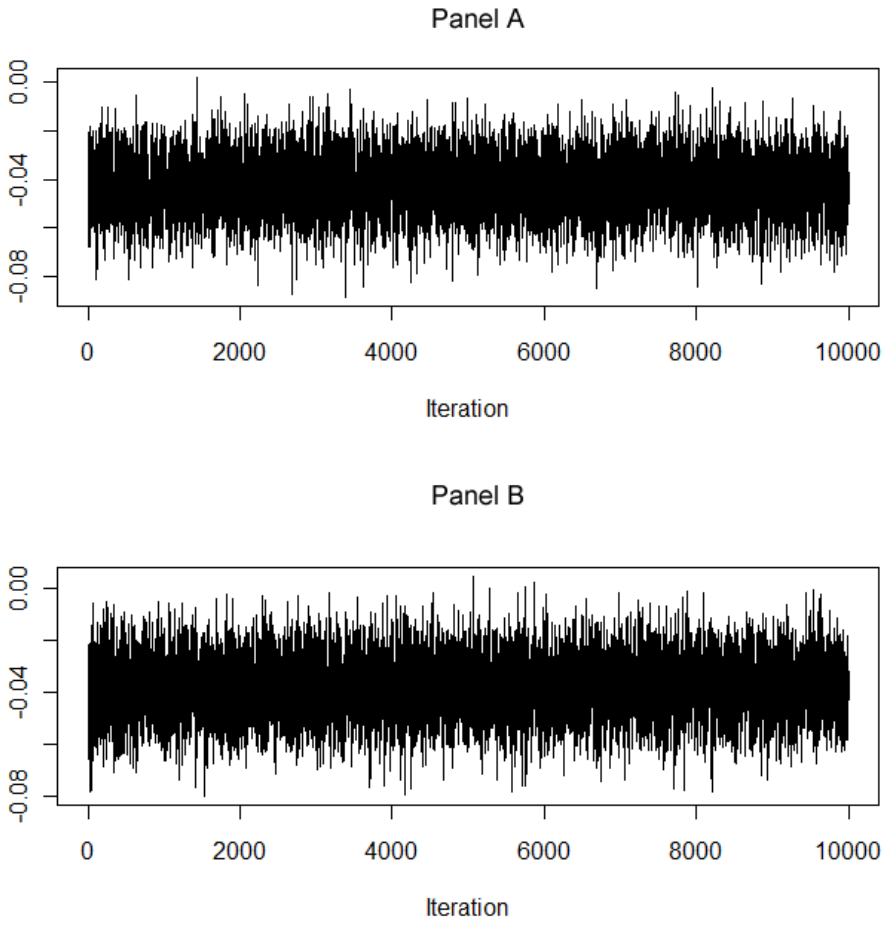

Figure 3.2: Trace Plots for County 1460. Panel A is the estimated random effect for 10,000 MCMC samples, after burn-in, using all data and the first set of initial conditions. Panel B is the estimated random effect for 10,000 MCMC samples, after burn-in, using all data and the second set of initial conditions.

"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 opts into or opts out of the data set" (Dwork & Lei, 2009). 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.

$R-U$ 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.

$R-U$ 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 3.3 and 3.4 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.

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 (3111) as compared to industry β (20), it may be reasonable to protect the estimates of u with more privacy (lower ϵ). Figures 3.5 and 3.6 show u having 10% and 88% of the plotted value of ϵ, respectively, while β accounts for

the remainder. For example, in Figure 3.5, 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 3.7 and 3.8 show u having 1% and 97% of the plotted privacy value of ϵ, respectively. Noticeable degradation is seen in Figure 3.7 when u is highly protected. For all figures except Figure 3.4, the privacy budget of the time trend was kept at 2%.

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 3.4.

Results from the analysis of deleting influential observations from Section V 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 was 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 correspond 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.

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 3.9 and 3.10 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.

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

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

$$u_s \sim N(0, \sigma_s^2 I), u_c \sim N(0, \sigma_c^2 I_{3111})$$

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

$$G = \begin{bmatrix} \sigma_s^2 I & 0 \\ 0 & \sigma_c^2 I_{3111} \end{bmatrix}$$

3.6.2 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 V-B.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 V-B.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

V-B.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 3.11 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 these 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 3.12.

3.7 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 Lapla-

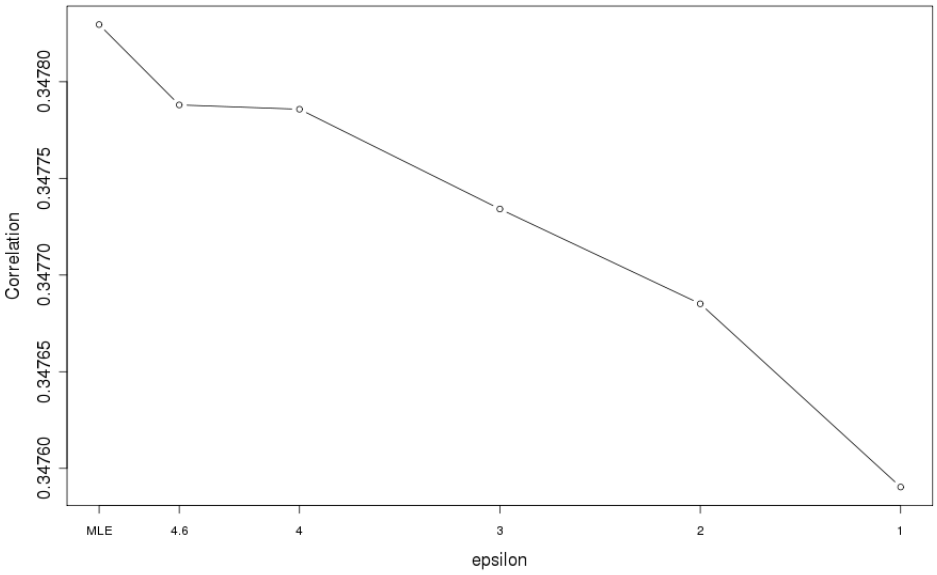

Figure 3.3: R-U Curve for JCR Linear Mixed Model with 49% ϵ budget for β and 49% for u

cian 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 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

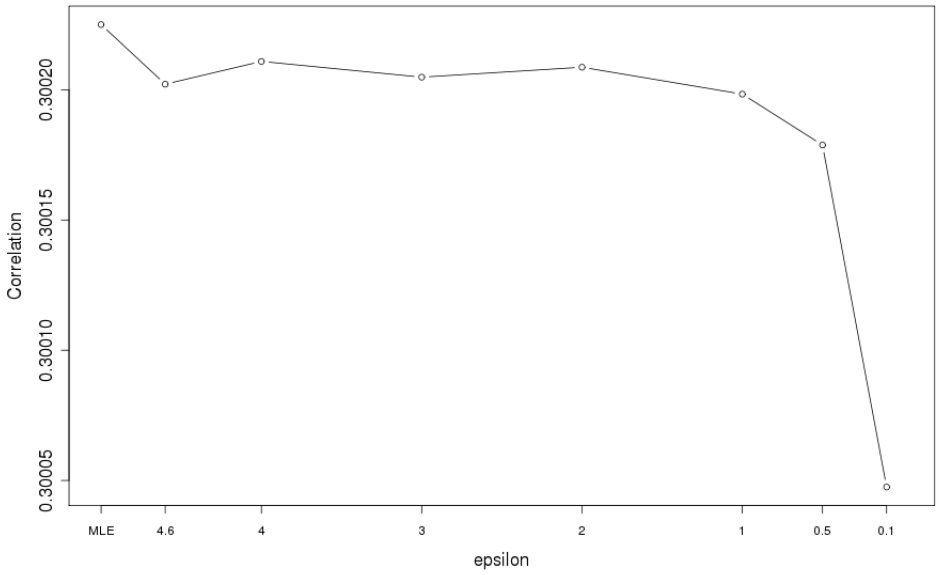

Figure 3.4: R-U Curve for JCR Linear Model with 100% ϵ budget for β

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.

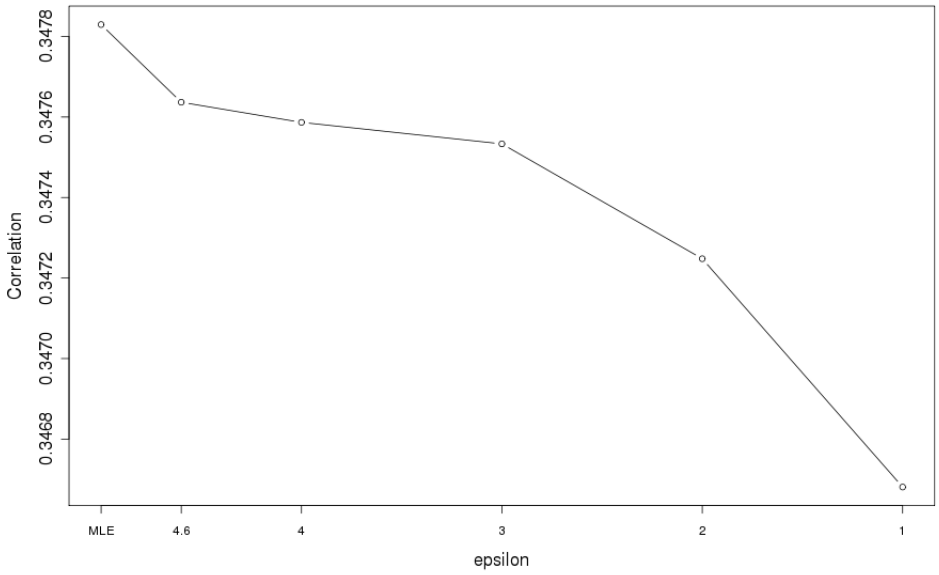

Figure 3.5: R-U Curve for JCR Linear Mixed Model with 88% ϵ budget for β and 10% for u

3.8 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 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

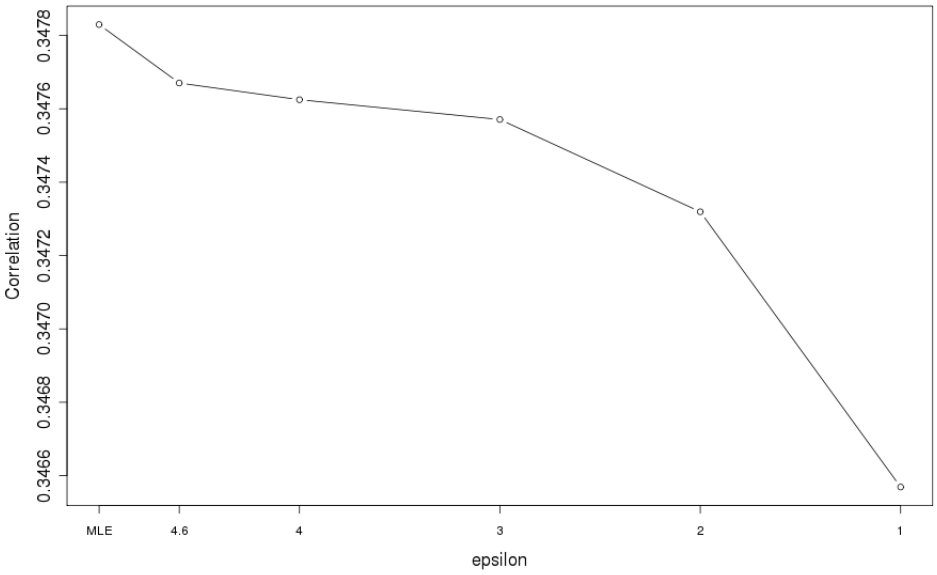

Figure 3.6: R-U Curve for JCR Linear Mixed Model with 10% ϵ budget for β and 88% for u

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

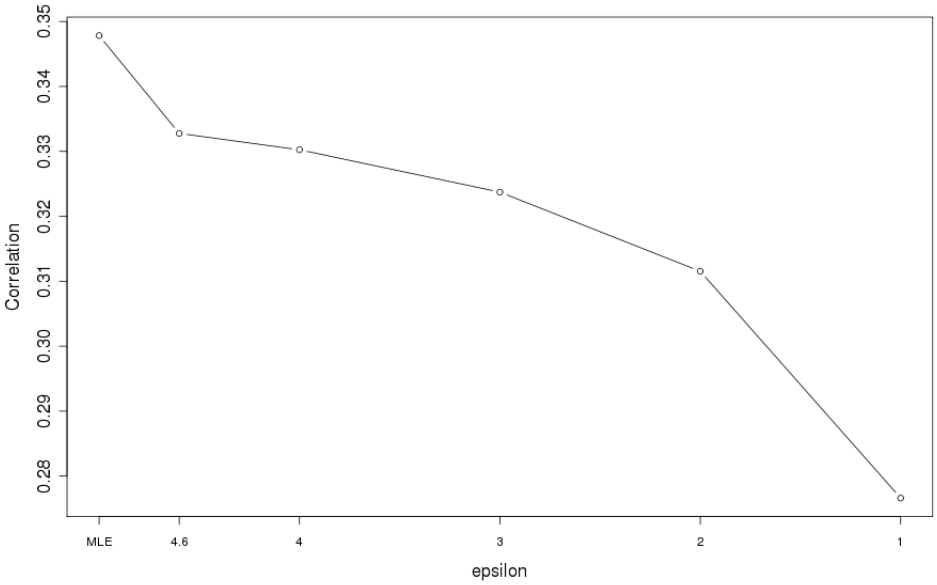

Figure 3.7: R-U Curve for JCR Linear Mixed Model with 97% ϵ budget for β and 1% for u

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.

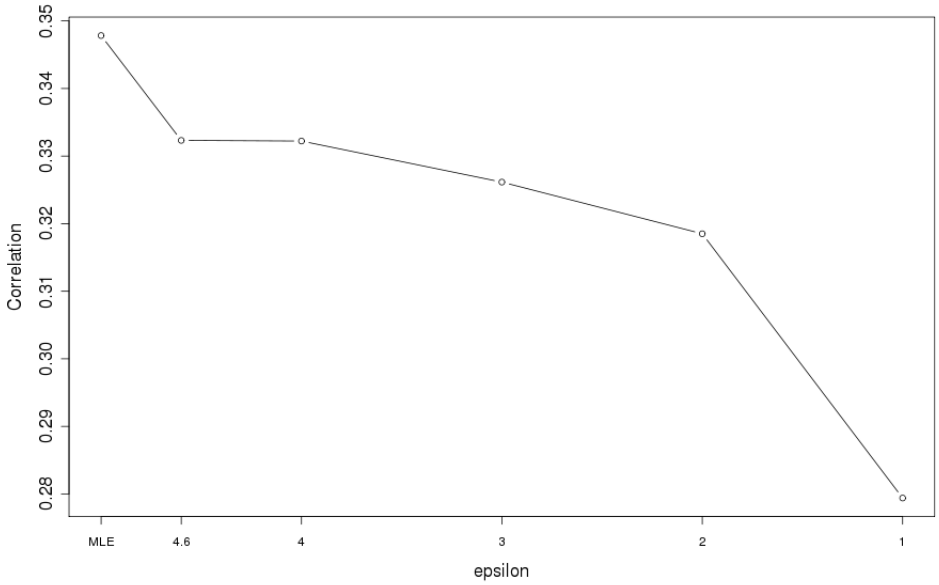

Figure 3.8: R-U Curve for JCR Linear Mixed Model with 1% ϵ budget for β and 97% for u

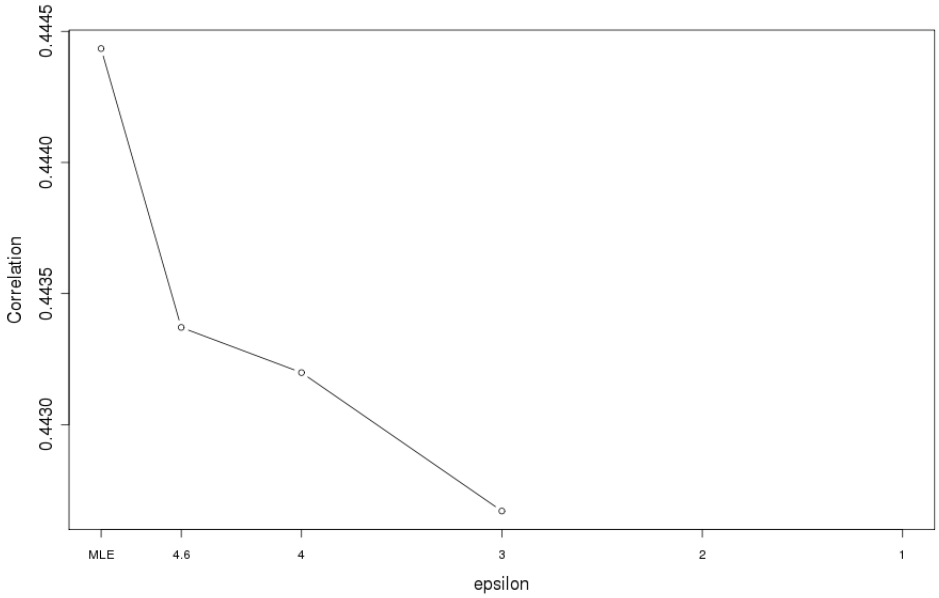

Figure 3.9: R-U Curve for JCR Linear Mixed Model with 88% ϵ budget for β and 5% for u and 5% for interactions

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 3.1: Technical Definitions

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
$\hat{\sigma}^2_\xi$	1	Residual Variance
$\hat{\sigma}^2_c$	1	County Variance

Table 3.2: Estimate Descriptions

Estimate	JCR Max Range	JCR Assumed Range	AR Max Range
$\hat{\beta}_1,...,\hat{\beta}_{20}$	2.490	2.00	49.60
$\hat{\beta}_{21}$	0.003	0.01	0.15
$\hat{u}_1,...,\hat{u}_{3111}$	2.040	2.00	374.50
$\hat{\sigma}^2_\xi$	0.200	0.50	5.50
$\hat{\sigma}^2_c$	0.190	0.50	24.10

Table 3.3: Maximum Empirical Ranges

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 3.4: Variance Estimate Ranges from Leaving out One County

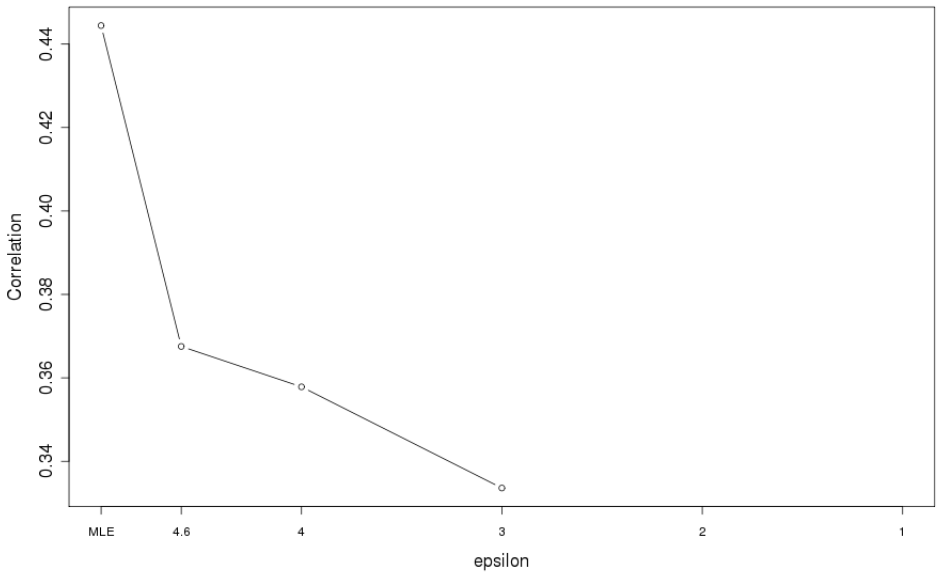

Figure 3.10: R-U Curve for JCR Linear Mixed Model with 97% ϵ budget for β and 0.5% for u and 0.5% for interactions

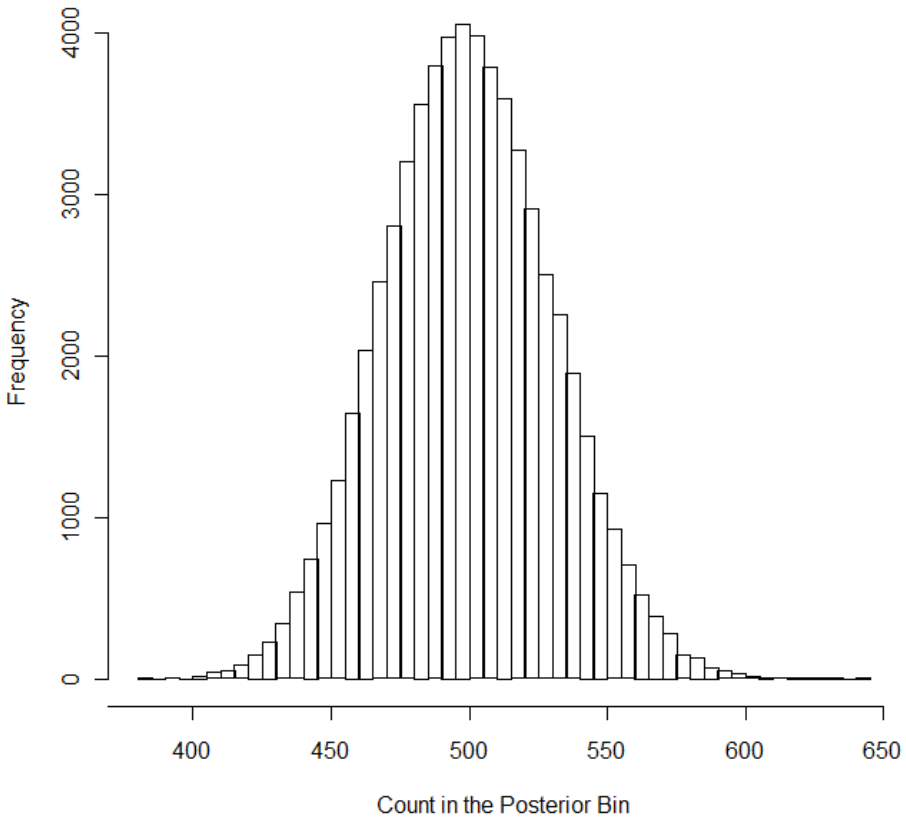

Figure 3.11: Histogram for the Replicated Model Including All Observations and Counties

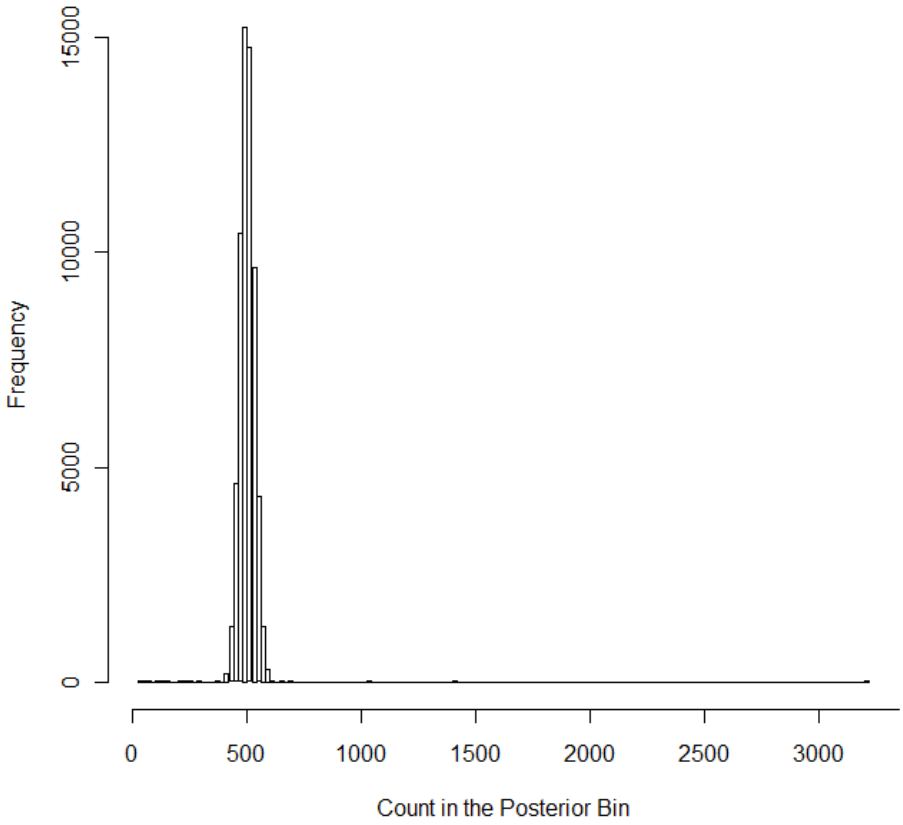

Figure 3.12: Histogram for the Model Deleting an Observation from County 3047

APPENDIX A

APPENDIX FOR PAPER 1

A.1 Standardizing forecasted change of M3-competition time series for ROC analysis

A limitation of the M3-competition data is that we have only one one-step-ahead forecast from each method for each series considered and cannot directly estimate the mean and variance of forecasts by method and series. Assume, however, that a method's forecasts have the same mean as a series (is an unbiased forecast) but have variance proportional to the data dependent on forecast method, but with proportion fixed across series for same forecast method. Consider the following notation:

Cross section of actual time series: $Y_{it}(i = 1, \ldots I; t = 1, \ldots T + m)$ where T= single, fixed forecast origin of M-Competition and m = forecast horizon (here we use $m = 1$ only)

Set of alternative method $j = 1, \ldots, J$ forecasts: $F_{ijt}(i = 1, \ldots I; j = 1, \ldots J; t = T + m)$

Sample statistics: Mean M_{it} and Variance S_{it} 1

Standardized actuals: $Y\prime_{it} = (Y_{it} - M_{it})/sqrt(S_{it})$

Decision thresholds U and L (fixed across time series):

If $Y\prime_{it} \geq U$ then high test positive (signaling a large increase)

If $Y\prime_{it} \leq L$ then low test positive (signaling a large decrease)

if $L < Y\prime_{it} < U$ then negative test

Assume that time series are stationary and that the mean of forecast method j for series i is M_{it}(independent of forecast method) but that the standard deviation is $k_j sqrt(S_{it})$. Thus we have:

Standardized forecasts: $F\prime_{ijt} = (F_{ijt} - M_{it})/k_j sqrt(S_{it})$

Decision Rules with control limits u and l (same across series and forecast methods because of standardization):

If $F\prime_{ijT+1} = (F_{ijT+1} - M_{it})/k_j sqrt(S_{it}) \geq U$ then high test positive or

If $F\prime_{ijT+1} = (F_{ijT+1} - M_{it})/sqrt(S_{it}) \geq k_j U$ then high test positive (use this rule for all series)

If $F\prime_{ijT+1} = (F_{ijT+1} - M_{it})/k_j sqrt(S_{it}) \leq L$ then low test positive or

If $F\prime_{ijT+1} = (F_{ijT+1} - M_{it})/sqrt(S_{it}) \leq k_j L$ then low test positive (use this rule for all series)

Take the case of forecast method j and low test positives. This is the key point: the right-hand value, $k_j L$, is of no concern because we process over values $min[(F_{ijT+1} - M_{it})/sqrt(S_{it})]$ to $max[(F_{ijT+1} - M_{it})/sqrt(S_{it})]$ with a grid to create the ROC curve for forecast method j. Regardless of the value of k_j, the ROC methodology produces a valid ROC curve and threshold values for the false positive rate. In other words, we do not depend on specific values for the right-hand side, but instead enumerate a grid of possible values for it depending on the sample of left-hand side values, from the smallest to largest possible

threshold values that do not result in constant decisions (all test positive or all test negative).

BIBLIOGRAPHY

Abowd, J., Stephens, B., Vilhuber, L. Andersson, F., McKinney, K., Roemer, M. & Woodcock, S. (2009). The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. in T. Dunne, J. Bradford, & M. Roberts, eds., Producer Dynamics: New Evidence from Micro Data, Chicago: University of Chicago Press for the NBER, 149-230.

Abowd, J., & Vilhuber, L. (2011). National Estimates of Gross Employment and Job Flows from the Quarterly Workforce Indicators with Demographic and Industry Detail. Journal of Econometrics, Vol. 161, 82-99.

Bates D. (2004). Sparse Matrix Representations of Linear Mixed Models. R Development Core Team.

Bates D. & Debroy S. (2004). Linear mixed models and penalized least squares. Journal of Multivariate Analysis: Vol 91: Iss. 1, 1-17.

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

Braga, A.A., Papachristos, A.V., & Hureau, D.M. (2012). The Effects of Hot Spots Policing on Crime: An Updated Systematic Review and Meta-Analysis. Justice Quarterly ahead-of-print, 1–31.

Chaudhuri, K., Monteleoni, C. & Sarwate, A. (2011). Differentially Private Empirical Risk Minimization. Journal of Machine Learning Research: 12, 1069-1109.

Clemen, R. T. (1989). "Combining Forecasts: A Review and an Annotated

Bibliography," International Journal of Forecasting, 5(4), 559–583.

Cohen, J., Garman, S., & Gorr, W. L. (2009). Empirical calibration of time series monitoring methods using receiver operating characteristic curves. International Journal of Forecasting, 25, 484–497.

Cohen, J., Gorr, W.L., & A. Olligschlaeger, A.M. (2007). Leading indicators and spatial interactions: a crime forecasting model for proactive police deployment. Geographical Analysis, 2007, 105–127.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. Statistical Review 15, 429-444.

Debroy, S. & Bates D. (2003). Computational Methods for Single Level Linear Mixed-effects Models. Technical Report No. 1073, Department of Statistics, University of Wisconsin.

DeLong, E. R., DeLong, D.M., & Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating curves: A nonparametric approach. Biometrics 44, 837-845.

Dorfman, D.D. & Alf, E. Jr. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals - Rating method data. J. Math. Psychol. 6, 487-496.

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

Dwork C. & Lei J. (2009). Differential Privacy and Robust Statistics. STOC.

Dwork C. & Smith A. (2009). Differential Privacy for Statistics: What we

Know and What we Want to Learn. Journal of Privacy and Confidentiality: Vol. 1: Iss. 2, Article 2.

Fawcett, T. (2006). An introduction to ROC Analysis. Pattern Recognition Letters, 27, 861–874.

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

Gorr, W. L. (2009). Forecast accuracy measures for exception reporting using receiver operating characteristics curves. International Journal of Forecasting, 25, 48–61.

Gorr, W.L. & Schneider, M.J. (2013). Large-change forecast accuracy: Reanalysis of M3-Competition data using receiver operating characteristic analysis. International Journal of Forecasting 29, 274–281.

Hadfield, J. (2010). MCMC Methods for MultiResponse Generalized Linear Mixed Models: The MCMCglmm R Packages. Journal of Statistical Software, 33(2), 1-22. URL http://www.jstatsoft.org/v33/i02/.

Hadfield, J. (2012). MCMCglmm Course Notes. Comprehensive R Archive Network, URL http://cran.rproject.org/web/packages/MCMCglmm/vignettes/CourseNote

Hanley, J. A., & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 839-843.

Hanley, J.A. & McNeil, B.J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology, 148, 839-843.

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

Hyndman, R. J. (2011), Data from the M-competitions, Comprehensive R Archive Network, http://cran.r-project.org/web/packages/Mcomp/Mcomp.pdf.

Hsu, M. J., & Hsueh, H. M. (2013). The linear combinations of biomarkers which maximize the partial area under the ROC curves. Computational Statistics, 1-20.

Janes, H., Longton, G. M., & Pepe, M.S. (2009). Accommodating covariates in ROC analysis. Stata J. 2009 January 1; 9(1): 17–39.

Komori, O., & Eguchi, S. (2010). A boosting method for maximizing the partial area under the ROC curve. BMC bioinformatics, 11(1), 314.

Koning, A. J., Franses, P.K., Hibon, M. & Stekler, H.O. (2005). The M3 competition: statistical tests of the results. International Journal of Forecasting 21, 397-409.

Kremer, M., Moritz, B., & Siemsen, E. (2011). Demand forecasting behavior: System neglect and change detection. Management Science, 57(10), 1827-1843.

Makridakis, S., Hibon, M. (2000), The M3-competition: results, conclusions and implications. International Journal of Forecasting, 16, 451-76.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008). Privacy: Theory Meets Practice on the Map. ICDE, 277-286.

McCulloch, C. & Searle, S. (2001). Generalized, Linear, and Mixed Models,

New York, NY: John Wiley and Sons, Inc..

Nissim, K. , Raskhodnikova, S., & Smith, A. (2007). Smooth Sensitivity and Sampling in Private Data Analysis. STOC.

Nobre, J.A. & Singer, J.d.M. (2007). Residual Analysis for Linear Mixed Models. Biometrical Journal 49: 6, 863-875.

Parker, C. (2011). An analysis of performance measures for binary classifiers. In Data Mining (ICDM), 2011 IEEE 11th International Conference on (pp. 517-526). IEEE.

Pepe, M. S., Cai, T., & Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. Biometrics, 62(1), 221-229.

Pinheiro, J. & Bates, D. (2000). Mixed Effects Models in S and S-Plus, New York, NY: Springer-Verlag New York, Inc.

Prasad, N. & Rao, J. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. Journal of the American Statistical Association: Vol. 85: No. 409, 163-171.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rao, J. (2003). Small Area Estimation, Hoboken, NJ: John Wiley and Sons, Inc.

Ricamato, M. T., & Tortorella, F. (2010). Combination of dichotomizers for

maximizing the partial area under the ROC curve. In Structural, Syntactic, and Statistical Pattern Recognition (pp. 660-669). Springer Berlin Heidelberg.

Ricketts, J.A. (1987). Management-by-exception reporting: An empirical investigation. Information & Management, 12, 235–246.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics, 12(1), 77.

Schabenberger, O. (2004). Mixed Model Influence Diagnostics. SUGI 29 Proceedings, Paper 189-29.

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

Smith, A. (2008). Efficient, Differentially Private Point Estimators. Preprint arXiv:0809.4794v1.

Snyder, R.D. & Koehler, A.B. (2008). Incorporating a tracking signal into a state space model. International Journal of Forecasting 25, 526-530.

Su, J. Q., & Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. Journal of the American Statistical Association, 88(424), 1350-1355.

Taylor, F.W. (1911). Shop management. Project Gutenberg eBook. Available-from: http://www.gutenberg.org/dirs/etext04/shpmg10.txt.

Wang, Z., & Chang, Y. C. I. (2011). Marker selection via maximizing the partial area under the ROC curve of linear risk scores. Biostatistics, 12(2), 369-385.

West, M, Harrison, P.J., & Mignon, H.S. (1985). Dynamic generalized linear models and Bayesian forecasting. Journal of the American Statistical Association, 80, 73–83.