



## Sufficiency Revisited: Rethinking Statistical Algorithms in the Big Data Era

Jarod Y. L. Lee, James J. Brown & Louise M. Ryan

To cite this article: Jarod Y. L. Lee, James J. Brown & Louise M. Ryan (2017) Sufficiency Revisited: Rethinking Statistical Algorithms in the Big Data Era, The American Statistician, 71:3, 202-208, DOI: [10.1080/00031305.2016.1255659](https://doi.org/10.1080/00031305.2016.1255659)

To link to this article: <https://doi.org/10.1080/00031305.2016.1255659>



View supplementary material [↗](#)



Accepted author version posted online: 15 Dec 2016.  
Published online: 18 Oct 2017.



Submit your article to this journal [↗](#)



Article views: 1751



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 6 View citing articles [↗](#)



## Sufficiency Revisited: Rethinking Statistical Algorithms in the Big Data Era

Jarod Y. L. Lee<sup>a,b</sup>, James J. Brown<sup>a,b</sup>, and Louise M. Ryan<sup>a,b,c</sup>

<sup>a</sup>School of Mathematical and Physical Sciences, University of Technology Sydney, Ultimo, NSW, Australia; <sup>b</sup>Australian Research Council Centre of Excellence for Mathematical & Statistical Frontiers, The University of Melbourne, Parkville, VIC, Australia; <sup>c</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

### ABSTRACT

The big data era demands new statistical analysis paradigms, since traditional methods often break down when datasets are too large to fit on a single desktop computer. Divide and Recombine (D&R) is becoming a popular approach for big data analysis, where results are combined over subanalyses performed in separate data subsets. In this article, we consider situations where unit record data cannot be made available by data custodians due to privacy concerns, and explore the concept of statistical sufficiency and summary statistics for model fitting. The resulting approach represents a type of D&R strategy, which we refer to as *summary statistics D&R*; as opposed to the standard approach, which we refer to as *horizontal D&R*. We demonstrate the concept via an extended Gamma–Poisson model, where summary statistics are extracted from different databases and incorporated directly into the fitting algorithm without having to combine unit record data. By exploiting the natural hierarchy of data, our approach has major benefits in terms of privacy protection. Incorporating the proposed modelling framework into data extraction tools such as TableBuilder by the Australian Bureau of Statistics allows for potential analysis at a finer geographical level, which we illustrate with a multilevel analysis of the Australian unemployment data. Supplementary materials for this article are available online.

### ARTICLE HISTORY

Received August 2015  
Revised October 2016

### KEYWORDS

Big data; Distributed database; Divide and recombine; Generalized linear mixed model; Multilevel model; Privacy



### 1. Introduction


The advent of big data has created a new research paradigm, with increasing reliance on large-scale administrative data from both public and private sectors (Einav and Levin 2014). These changes have had concomitant impact on statistical analysis. The traditional practice of performing statistical analysis using a single combined dataset is often infeasible due to memory and storage limitations of standard computers. Adding to these issues are privacy concerns, which often render data custodians reluctant to release unit record data. These issues combine to limit the ability of analysts to fully unlock the actionable information in big data.

As a solution to the memory and storage limitations problem, Divide and Recombine (D&R) has been proposed as an effective, generic approach to statistical analysis of big data (Guha et al. 2012). D&R involves (i) dividing data into manageable subsets, (ii) performing statistical analysis independently on each subset, and then (iii) combining the results, typically via some form of averaging (Figure 1). Data are typically divided via either *replicate* division or *conditioning variable* division (Bühlmann et al. 2016, chap. 3). Replicate division divides the data based on random sampling without replacement, whereas conditioning variable division stratifies the data according to one or more variables in the data. An example of conditioning variable division is to partition disease incidence data by postal areas. The results are combined in a way that results in the least discrepancy

compared to the *all data estimate*, that is, estimate obtained by using the entire dataset. Except in very simple cases, D&R results are approximate.

DeltaRho (formerly Tessera) (Bühlmann et al. 2016, chap. 3) is an open source implementation of D&R that combines the R statistical programming environment (R Core Team 2014) at the front end with various back end options such as Hadoop (White 2009) and Spark (Zaharia et al. 2010, 2012). This allows users to scalably leverage all of the statistical methods readily available in R while abstracting the technical programming details, making D&R more accessible to the general statistical community. The emergence of these systems has sparked research interest in the D&R algorithm. A selection of recent examples include Boyd et al. (2011); Chu, Keshavarz, and Boyd (2013); Lubell-Doughtie and Sondag (2013); Scott et al. (2016); Chen and Xie (2014); Kleiner et al. (2014); Minsker et al. (2014); Neiswanger, Wang, and Xing (2014); Xu et al. (2014); Perry (2017); and Miroshnikov, Wei, and Conlon (2015). In typical D&R applications, we have unit record data where analysis on a single machine is not feasible, because the data are either too large to store, or of moderate size but the statistical method being used is too computationally intensive. The dataset is divided into subsets of similar structure and the intended analysis is performed on each of the subsets. We refer to this kind of division as *horizontal D&R*, where unit level data are partitioned in such a way that each subset holds the same variables but for different cases.

**CONTACT** Jarod Y. L. Lee  [yan.lee@uts.edu.au](mailto:yan.lee@uts.edu.au)  School of Mathematical and Physical Sciences, University of Technology Sydney, Ultimo, NSW 2007, Australia and Australian Research Council Centre of Excellence for Mathematical, & Statistical Frontiers, The University of Melbourne, Parkville, VIC 3010, Australia.

 Supplementary materials for this article are available online. Please go to [www.tandfonline.com/r/TAS](http://www.tandfonline.com/r/TAS).

© 2017 American Statistical Association

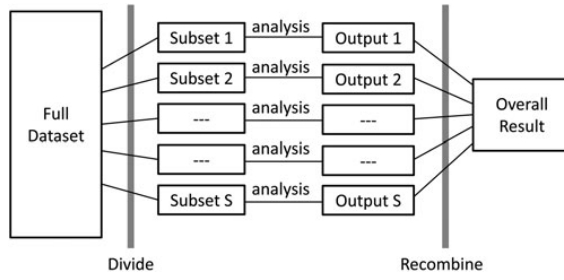


Figure 1. The Divide and Recombine (D&R) framework.

In this article, we consider situations where unit record data cannot be made available due to privacy reasons, even after personal identifiers such as name, address, date of birth, and ID number have been removed. This situation arises often in practice because the presence of rich information, when combined with the use of sophisticated data mining tools, renders privacy breaching a major threat (Fienberg 2006). This is true even after statistical disclosure control methods (Hundepool et al. 2012) have been applied to safeguard the confidentiality of data (Sweeney 2002; Coull et al. 2007; Homer et al. 2008; Narayanan and Shmatikov 2008).

As a solution, we explore the concept of statistical sufficiency and summary statistics for model fitting. Sufficiency is a concept taught in every introductory mathematical statistics course, but it has not been actively used for practical model fitting because the need has not been there. The aim is to compress the raw data in each subset into low-dimensional summary statistics for model fitting. We refer to this as *summary statistics D&R*, emphasizing the fact that unit record data cannot be made available, as opposed to horizontal D&R. We illustrate the concept via a multilevel model (Gelman and Hill 2006; Goldstein 2011) based on an extension of the Gamma–Poisson model by Christiansen and Morris (1997). In this context, the use of summary statistics exploits the natural grouping structure in the data and allows the direct modeling of data from multiple sources using summary information, without the need to combine them into a single file, thus is privacy preserving. We apply the model to publicly available unemployment data from the Australian Bureau of Statistics and explain the benefit of our framework in terms of allowing analysis at a finer geographical level.

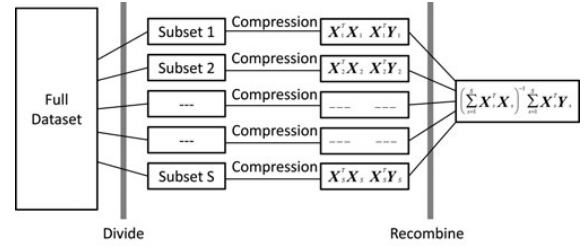
The article is organized as follows. In Section 2, we motivate the distinction between summary statistics D&R and horizontal D&R using simple linear regression as an example. We then describe the proposed extended Gamma–Poisson model in Section 3. In Section 4, the model is applied to the Australian unemployment data. We close with some concluding remarks in Section 5.

## 2. Illustrative Example

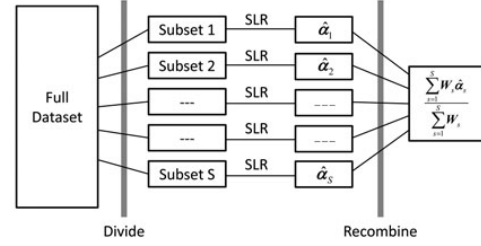
A linear regression takes the general form

$$Y = X\alpha + \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

where  $Y$  is a  $n \times 1$  vector of response variables,  $X$  is a  $n \times p$  design matrix,  $\alpha$  is a  $p \times 1$  vector of regression parameter,



(a)



(b)

Figure 2. Linear regression via Divide and Recombine (D&R). (a) Summary statistics D&R; (b) Horizontal D&R.

$\epsilon$  is a  $n \times 1$  vector of independent errors, and  $\Sigma$  is a  $n \times n$  diagonal matrix, with common diagonal elements  $\sigma^2$ . Standard least squares and maximum likelihood estimates give the *all data estimate*  $\hat{\alpha} = (X^T X)^{-1} X^T Y$ . When data are too large to fit into a single machine, we can resort to two different approaches: (i) summary statistics D&R and (ii) horizontal D&R.

Summary statistics D&R (Figure 2(a)) includes the following steps: (i) divide the data into  $S$  subsets of similar structure, with  $Y_s$  denoting the vector of responses in subset  $s$  and  $X_s$  the corresponding design matrix, (ii) calculate two sets of summary data for each of the subsets, that is,  $X_s^T X_s$  and  $X_s^T Y_s$ , then (iii) combine via  $(\sum_{s=1}^S X_s^T X_s)^{-1} \sum_{s=1}^S X_s^T Y_s$ . Chen et al. (2006) referred to this as the *regression cube* technique. The resulting *aggregated estimate* is exactly equivalent to the all data estimate due to the matrix properties  $X^T X = \sum_{s=1}^S X_s^T X_s$  and  $X^T Y = \sum_{s=1}^S X_s^T Y_s$ .

Horizontal D&R (Figure 2(b)) includes the following steps: (i) divide the data into  $S$  subsets of similar structure, (ii) perform independent least-squares regression on each subset to obtain  $\hat{\alpha}_s = (X_s^T X_s)^{-1} X_s^T Y_s$  and then (iii) weight the results to obtain the aggregated estimate  $\sum_{s=1}^S W_s \hat{\alpha}_s / \sum_{s=1}^S W_s$ . The optimal weight is obtained using  $W_s = X_s^T X_s$ , which is proportional to the inverse variance-covariance matrix of the regression parameter. This results in an aggregated estimate that is exactly equivalent to the one obtained via the summary statistics approach. Intuitively, the use of inverse variance-covariance matrix as weight makes sense as we are giving larger credibility to subsets with lower variability.

Summary statistics D&R differs from horizontal D&R. For summary statistics D&R, we extract the relevant summary statistics that best summarizes data in each subset, so that the resulting aggregated estimate is as “close” as possible to the all data estimate. Only summary data are extracted, rendering unit record data unnecessary. For horizontal D&R, we perform the

intended statistical analysis (linear regression in this case) independently on each subset and choose an aggregate estimate to minimize the error. Unit record data are typically required so that the intended analysis can be done on each subset.

For linear regression, both aggregated estimates are exactly equivalent to the all data estimate due to the linearity of the estimating equation in  $\alpha$ , assuming optimal weights have been used. This is generally not true for more complicated models such as logistic regression (Xi, Lin, and Chen 2009) and nonlinear estimating equations (Lin and Xi 2011), where we can only hope to find aggregated estimators that are consistent.

### 3. Extended Gamma–Poisson Model

In this section, we propose a model for correlated data that characterizes individual level event rates as a function of both individual and area level covariates, and show that it can be fitted using sufficient and summary statistics. Our approach extends the Gamma–Poisson model by Christiansen and Morris (1997) to include both individual and area level predictors. The Poisson component characterizes the effect of individual level variables on the event rate. The Gamma component incorporates the effects of area level covariates as well as a random component that reflects area to area variation that is not captured by area level covariates.

#### 3.1. The Model

Let  $Y_{ij}$  denotes whether or not the  $i$ th individual in the  $j$ th area experienced the event of interest and let  $\mathbf{x}_{ij}$  be the  $p \times 1$  vector of covariates measured on this individual, with  $x_{ij,1} = 1$  to allow for the intercept. Given an area specific random effect  $b_j$ , we assume that  $Y_{ij}$  is independently Poisson distributed with mean  $\lambda_{ij} = b_j \exp(\mathbf{x}_{ij}^T \alpha)$ , where  $\alpha$  is a  $p \times 1$  vector of unknown regression coefficients to be estimated. The random effect  $b_j$  is assumed to follow a Gamma distribution with parameters chosen so that the mean  $\mu_j = \exp(\mathbf{u}_j^T \gamma)$  and the variance is  $\kappa \mu_j$ , where  $\mathbf{u}_j$  is a vector of area level covariates,  $\gamma$  is the corresponding vector of unknown regression parameters to be estimated, and  $\kappa$  is a dispersion parameter to be estimated. That is,

$$b_j \stackrel{\text{ind.}}{\sim} \text{Gamma}\left(\frac{\mu_j}{\kappa}, \kappa\right) \\ \stackrel{\text{ind.}}{\sim} \text{Gamma}[\mu_j, \kappa \mu_j].$$

The random effect  $b_j$  captures the deviation of the area specific rates from the mean outcome, taking into account area level variables as well as any remaining unexplained variation. In our formulation,  $b_j > 1$  corresponds to a positive deviation of area  $j$  from the mean rate, whereas  $0 < b_j < 1$  corresponds to a negative deviation. Note that we parameterize the Gamma distribution in terms of an area specific mean parameter  $\mu_j > 0$  and a constant scale parameter  $\kappa > 0$ . The round parentheses indicate the standard shape and scale parameterization of the Gamma distribution, whereas the square brackets indicate the mean and variance formulation.

We choose to model a binary response variable using a Poisson model, as the Poisson distribution is a good approximation to the Binomial distribution when dealing with relatively rare events (events where the chance of a success on any particular trial is small) such as unemployment and heart disease. The Poisson model is often preferred because the covariate effects can be directly interpreted as risk ratios due to the log canonical link. Although, we focus on the case where  $Y_{ij}$  is a binary 0 or 1 variable, extension to the more general case where  $Y_{ij}$  can be any integer counts is straightforward.

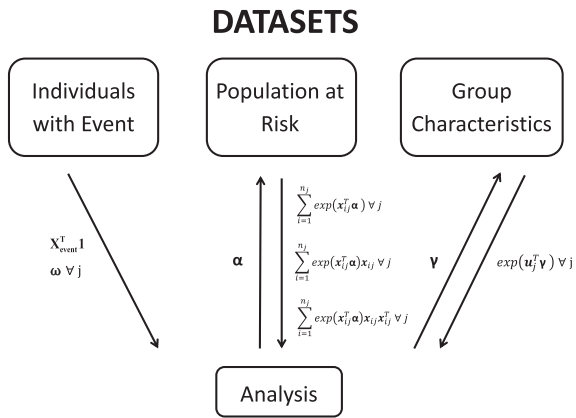
#### 3.2. Model Fitting Using Summary Data

With some straightforward algebra, the log-likelihood equation  $\ell(\alpha, \gamma, \kappa; \mathbf{y})$  can be written as

$$\ell(\alpha, \gamma, \kappa; \mathbf{y}) = \sum_j \left( \sum_{i:Y_{ij}=1} \mathbf{x}_{ij}^T \alpha \right) + \sum_j \left\{ -\frac{\mu_j}{\kappa} \log(\kappa) - \log \Gamma\left(\frac{\mu_j}{\kappa}\right) + \log \Gamma\left(\omega_j + \frac{\mu_j}{\kappa}\right) - \left(\omega_j + \frac{\mu_j}{\kappa}\right) \log \left[ \sum_{i=1}^{n_j} e^{\mathbf{x}_{ij}^T \alpha} + \frac{1}{\kappa} \right] \right\}.$$

The log-likelihood is a function of a few sufficient and summary statistics from various databases. Figure 3 presents a summary diagram of the data requirement to obtain the using maximum likelihood estimates via the Newton–Raphson algorithm. For individuals who experienced the event, the summary data required are:

1. For categorical variables, the number of people who had the event within every level of individual level variables of interest; for continuous variables, the sum of individual level variables of interest across individuals who experienced the event. For the gender (categorical) and age (continuous) variables, this translates into the number of events among either males or females and the sum



**Figure 3.** Data requirement to obtain the maximum likelihood estimates of the extended Gamma-Poisson model via the Newton–Raphson algorithm. Only a few sufficient and summary statistics are required without needing to access the unit record data.

of ages for all individuals who experienced the event. Technically, we write this as  $\mathbf{X}_{\text{event}}^T \mathbf{1}$ , where  $\mathbf{X}_{\text{event}}$  is a  $M \times p$  matrix where each row comprises a row vector of covariates for one of the  $M$  individuals who experienced the event of interest. Hence,  $\mathbf{X}_{\text{event}}^T \mathbf{1}$  is simply the column totals from  $\mathbf{X}_{\text{event}}$ .

2. The number of subjects who had the event in each area, defined as  $\omega_j$  for area  $j$ .

The summary data above only need to be computed once, since they do not depend on any unknown model parameters being estimated. Technically, these are sufficient statistics (Casella and Berger 2002). For the dataset on the population at risk, three sets of summary statistics are required from each area  $j$ . These involve summations over all individuals living in the common area, rendering individual level data unnecessary. These are not sufficient statistics since they involve the unknown parameter  $\alpha$ . A similar process applies for the area dataset. At each iteration of the algorithm, likelihood contributions are computed as functions of these summary statistics, leading to an improved value of the unknown parameters. The process repeats until convergence.

As an illustration, the framework can be applied to hospital variation studies whose goal is to quantify variation in hospital admission rates as a function of a variety of individual level factors such as age, gender, medical history, as well as hospital level factors such as proportion of interns, ratio of residents to beds, hospital resources, and area level socio-economic advantage. Patients' data are confidential and hospitals are obliged to protect them. De-identifying individual level data is not sufficient to prevent disclosure, and analysts who wish to obtain the data have to go through an ethics application, which is time consuming if not impossible. However, hospitals may be quite willing to provide the summary statistics required by the proposed framework. These summaries are then passed to an analysis computer via a network, which returns an improved value of the unknown parameters and passed back to the hospitals to compute a new set of summary statistics. The process iterates until convergence.

### 3.3. Parameter Initialization

A good starting value is essential to ensure proper convergence. Here, we propose an initialization process for the regression coefficients using summary data. Assuming an independent structure, the parameters of standard log-linear Poisson models can be estimated via the Newton-Raphson algorithm

$$\alpha \leftarrow \alpha + (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{Y} - \mu),$$

where  $\mathbf{A} = \text{diag}(\text{var}(Y_{ij}) d\eta_{ij}/d\mu_{ij})$  and  $\mathbf{W} = \mathbf{A}(d\eta_{ij}/d\mu_{ij})^{-1}$ , with  $\eta$  being the link function.

For log-linear models where the outcome is either a success or a failure,  $\mathbf{A} = \text{diag}(1)$  and  $Y_i = 0, 1$ . Thus, the equation above can be rewritten as

$$\alpha \leftarrow \alpha + \left( \mathbf{X}_{\text{pop}}^T \mathbf{W} \mathbf{X}_{\text{pop}} \right)^{-1} \left( \mathbf{X}_{\text{event}}^T \mathbf{1} - \mathbf{X}_{\text{pop}}^T \mu \right),$$

where  $\mathbf{X}_{\text{event}}$  and  $\mathbf{X}_{\text{pop}}$  are the model design matrices for individuals with event and the entire population at risk respectively,

$\mu = \exp(\mathbf{X}_{\text{pop}} \alpha)$  and  $\mathbf{W} = \text{diag}(\mu)$ . Thus, we only require a set of sufficient statistics for individuals with event ( $\mathbf{X}_{\text{event}}^T \mathbf{1}$ ) and two sets of summary statistics for the population at risk for each area ( $\mathbf{X}_j^T \mathbf{W}_j \mathbf{X}_j \forall j$  and  $\mathbf{X}_j^T \mu_j \forall j$ ), due to the matrix properties  $\mathbf{X}^T \mathbf{W} \mathbf{X} = \sum_j \mathbf{X}_j^T \mathbf{W}_j \mathbf{X}_j$  and  $\mathbf{X}^T \mu = \sum_j \mathbf{X}_j^T \mu_j$ .

These sufficient and summary statistics coincide with those in Figure 3. In other words, the same set of sufficient and summary statistics can be used for parameter initialization and model fitting.

## 4. Application: Australian Unemployment Data

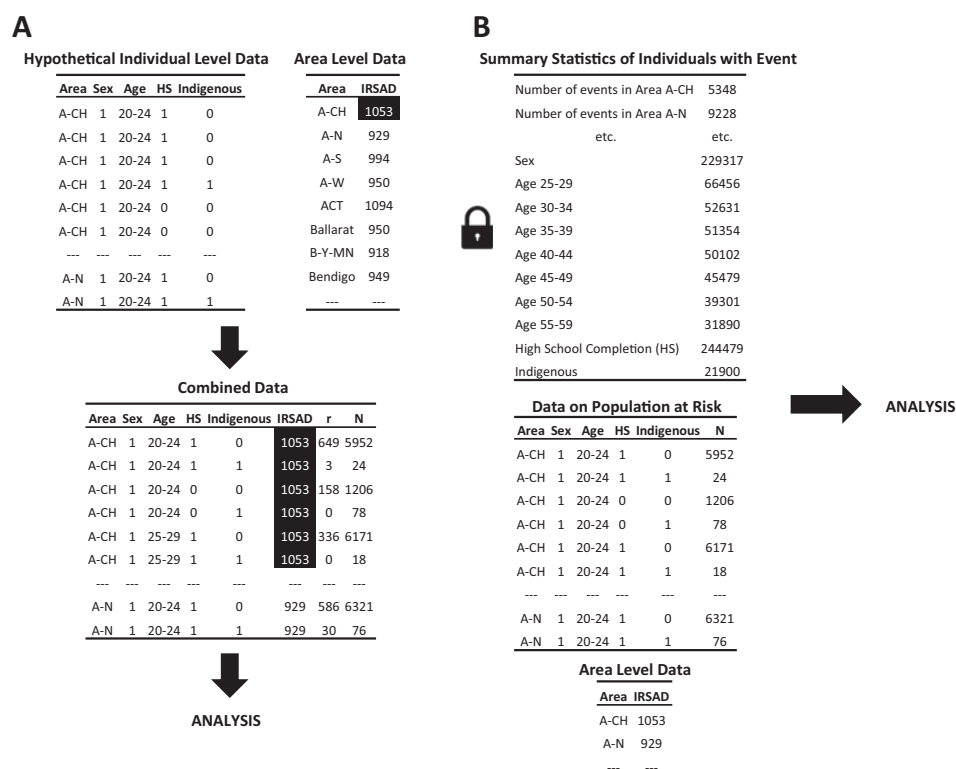
The dependence of the extended Gamma-Poisson model on only a few sufficient and summary statistics allows the fitting of detailed statistical models without actually having access to unit record data. This has important applications in terms of privacy protection in distributed databases, which we illustrated using the hypothetical hospital variation studies example in Section 3.2. This section aims to demonstrate the benefit of incorporating the proposed model into data extraction tools such as TableBuilder by the Australian Bureau of Statistics (ABS).

We obtain the Australian unemployment data from TableBuilder, an online tool by the ABS whereby users can build cross-classifications of census variables for geographical areas as defined in the Australian Statistical Geography Standard (ASGS; Australian Bureau of Statistics 2012). We wish to explore whether individuals living in areas in 2011 that had higher socio-economic advantage prior to the Global Financial Crisis (2007–2008) have more resilience to unemployment, adjusting for individual level variables such as sex, age, high school completion and identifying as an indigenous Australian. The individual level variables sex, age, high school completion, and indigenous status in 2011 are cross-classified according to Statistical Area Level 4 (SA4). SA4 is chosen in this context as it was originally designed for the outputs of the Australian Labour Force Survey. For a measure of socio-economic advantage, we use the Index of Relative Advantage and Disadvantage (IRSAD) (Australian Bureau of Statistics 2008) that can be obtained from the ABS in a separate database. High values of IRSAD indicate high social advantage, and vice versa. We use IRSAD values from the 2006 Census, the most recent Census preceding 2011.

We now have data on individuals with events and population at risk from TableBuilder, and data on area level covariates from a separate ABS database. Available packages for fitting multilevel models in standard statistical softwares such as lme4 (Bates et al. 2015) in R (R Core Team 2014) require data to be combined (Figure 4(a)). This results in unnecessary repetition of area level covariates over multiple rows (see the shaded cells in Figure 4(a)). With the extended Gamma-Poisson model, we can avoid the step of combining these data into a single large design matrix by computing just a few summary statistics directly from each dataset (Figure 4(b)).

Moreover, privacy legislation precludes anyone apart from ABS employees from having access to individual level census data. Only confidentialized tabular data can be made available to researchers, whereby the counts are randomly adjusted to





**Figure 4.** Data requirement. (a) Existing software packages for fitting multilevel models require the individual and area level data to be combined into a single file before performing analysis. This results in the unnecessary repetition of area level variables as indicated by the shaded cells. (b) By using the extended Gamma-Poisson model, datasets are analyzed directly without the need to combine them. In addition, for individuals who experienced the event, the model only requires sufficient statistics instead of the full dataset.

reduce the risk of disclosure (O’Keefe 2008; Leaver 2009). This means that we need to be cautious when using data from tables with small cell counts, since they are likely to be unreliable. This is especially true for the event counts (denoted by  $r$  in Figure 4(a)), even more when the event under consideration is rare or when there are many cross-classifications. In our application, it would be ideal to model at the finer SA3 level as it provides a more detailed analysis compared to the Australian Labor Force Survey. However, small cell counts prevents us from doing so. In this regard, extending tools such as TableBuilder to only output the sufficient statistics of individuals with event required by the Gamma-Poisson model would be useful (Figure 4(b)).

We fit the extended Gamma-Poisson model to the unemployment data, and compare the results with the Normal-Poisson model fitted on the combined data using the lme4 package (Bates et al. 2015) in R (R Core Team 2014). The estimates and standard errors produced by both models are very similar. The results, summarized in Table 1, reveal some interesting patterns. Unemployment rates are lower for males ( $p < 0.001$ ) and tend to decrease with age. The unemployment rate of a person who completed high school is 36.87% lower than for a person who did not complete high school ( $p < 0.001$ ). The unemployment rate of indigenous Australians is 115.98% higher than non-indigenous Australians, even after controlling for education ( $p < 0.001$ ), demonstrating the continued need to reduce economic disadvantage for this community. After adjusting for individual level characteristics,

**Table 1.** Estimates and standard errors based on the Australian unemployment data, fitted using the Normal-Poisson model and the extended Gamma-Poisson model. “Est” and “SE” correspond to the estimates and standard errors, respectively.

Parameter	Normal-Poisson		Gamma-Poisson	
	Est	SE	Est	SE
$\alpha_0$ (Intercept)	-2.06*	0.028	-2.03*	0.027
$\alpha_1$ (Female)			Reference Group	
$\alpha_1$ (Male)	-0.06*	0.003	-0.06*	0.003
$\alpha_2$ (age 20 to 24)			Reference Group	
$\alpha_2$ (age 25 to 29)	-0.51*	0.005	-0.51*	0.005
$\alpha_3$ (age 30 to 34)	-0.70*	0.005	-0.70*	0.005
$\alpha_4$ (age 35 to 39)	-0.80*	0.005	-0.80*	0.005
$\alpha_5$ (age 40 to 44)	-0.92*	0.006	-0.92*	0.006
$\alpha_6$ (age 45 to 49)	-1.03*	0.006	-1.03*	0.006
$\alpha_7$ (age 50 to 54)	-1.11*	0.006	-1.11*	0.006
$\alpha_8$ (age 55 to 59)	-1.10*	0.007	-1.10*	0.007
$\alpha_9$ (Not completed High School)			Reference Group	
$\alpha_9$ (Completed High School)	-0.46*	0.003	-0.46*	0.003
$\alpha_{10}$ (Not Indigenous)			Reference Group	
$\alpha_{10}$ (Indigenous)	0.77*	0.007	0.77*	0.007
$\gamma$ (IRSAD)	-0.04	0.027	-0.04	0.027
$\sigma^2$	0.07	N/A	N/A	N/A
$\kappa$	N/A	N/A	0.06*	0.010

NOTE: \* Significant at  $p < 0.001$ .

area level measures of social advantage have only a modest impact ( $p = 0.138$ ) on unemployment rates. However, there remains significant area-to-area variation in unemployment rates.

## 5. Concluding Remarks

The article argues that statistical sufficiency and summary statistics offer an attractive framework in the big data era, especially in the setting of large-scale administrative databases where privacy concerns prevent general access to unit record data. The concept is illustrated via an extended Gamma–Poisson multilevel model. The model works by gathering relevant pieces of summary information required for construction of the log-likelihood directly from the separate data sources. This is a natural solution since the relevant data are often drawn from different sources anyway. For example, epidemiologists often augment their study populations with information about the communities in which their study participants live. Such community-level variables might be obtained from a national census or from other surveys. As another example, information about the study population may come from different hospital administrative databases, held locally at the respective hospitals. Sharing these databases among hospitals might not be possible due to privacy reasons. Aside from offering benefits in terms of privacy protection, incorporating the model into data extraction tools such as TableBuilder allows for potential analysis at a finer geographical level.

The ideas discussed in this article bear some connection to symbolic data analysis (Billard and Diday 2006), where data are compressed into distributions such as hyperrectangles or histograms, rather than a single summary. The main difference is that in symbolic data analysis, exact statistical analysis is performed using approximate data (e.g., loss of information when summarizing individual data points into histograms); whereas in Gamma–Poisson model, exact analysis is performed using exact data (using summary statistics does not result in any loss of information under the Gamma–Poisson model, compared to using unit record data).

The proposed model has the potential to become a valuable addition to the statistician's toolbox in the quest to make better use of the ever increasing volumes of data being generated in the big data era (Einav and Levin 2014). More generally, the model and analysis we developed and implemented in this article are example of rethinking classic statistical ideas for model fitting in the big data era. There is great potential for developing new algorithms that can be used in the analysis of large administrative databases. For example, in the case of *vertical D&R* where data are partitioned in such a way that each partition hold a subset of the variables for the common individuals, there is still considerable methodological work to be done. It would be good to see more of these developments happening in the statistics literature.

## Supplementary Material

**Derivation of the extended Gamma–Poisson fitting algorithm:** Derivation of the log-likelihood, score vectors and Hessian matrices required for model fitting.

## Funding

This research was partially supported by the Australian Bureau of Statistics.

## References

- Australian Bureau of Statistics (2008), "Information Paper: An Introduction to Socio-Economic Indexes for Areas (SEIFA)," ABS Catalogue No. 2039.0, Australian Bureau of Statistics. [205]
- (2012), "Australian Statistical Geography Standard (ASGS): Correspondences," ABS Catalogue No. 1270.0.55.006, Australian Bureau of Statistics. [205]
- Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015), *lme4: Linear Mixed-Effects Models Using Eigen and S4*, R package version 1.1-8. Available at <http://CRAN.R-project.org/package=lme4> [205,206]
- Billard, L., and Diday, E. (2006), *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, New York: Wiley. [207]
- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, 3, 1–122. [202]
- Bühlmann, P., Drineas, P., Kane, M., and van der Laan, M. (eds.) (2016), *Handbook of Big Data*, Boca Raton, FL: CRC Press. [202]
- Casella, G., and Berger, R. L. (2002), *Statistical Inference*, Pacific Grove, CA: Duxbury. [205]
- Chen, X., and Xie, M. (2014), "A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data," *Statistica Sinica*, 24, 1655–1684. [202]
- Chen, Y., Dong, G., Han, J., Pei, J., Wah, B. W., and Wang, J. (2006), "Regression Cubes with Lossless Compression and Aggregation," *IEEE Transactions on Knowledge and Data Engineering*, 18, 1585–1599. [203]
- Christiansen, C., and Morris, C. (1997), "Hierarchical Poisson Regression Modeling," *Journal of the American Statistical Association*, 92, 618–632. [203,204]
- Chu, E., Keshavarz, A., and Boyd, S. (2013), "A Distributed Algorithm for Fitting Generalized Additive Models," *Optimization and Engineering*, 14, 213–224. [202]
- Coull, S., Collins, M., Wright, C., Monrose, F., and Reiter, M. (2007), "On Web Browsing Privacy in Anonymized NetFlows," in *Proceedings of 16th USENIX Security Symposium*, pp. 339–352. [203]
- Einav, L., and Levin, J. (2014), "Economics in the Age of Big Data," *Science*, 346, 479–480. [202,207]
- Fienberg, S. E. (2006), "Privacy and Confidentiality in an E-Commerce World: Data Mining, Data Warehousing, Matching and Disclosure Limitation," *Statistical Science*, 21, 143–154. [203]
- Gelman, A., and Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge, UK: Cambridge University Press. [203]
- Goldstein, H. (2011), *Multilevel Statistical Models*, New York: Wiley. [203]
- Guha, S., Hafen, R., Rounds, J., Xia, J., Li, J., Xi, B., and Cleveland, W. S. (2012), "Large Complex Data: Divide and Recombine (D&R) with RHIP," *Stat*, 1, 53–67. [202]
- Homer, N., Szlinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J., Stephan, D., Nelson, S., and Craig, D. (2008), "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures using High-Density SNP Genotyping Microarrays," *PLoS Genetics*, 4, [203]
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and De Wolf, P. P. (2012), *Statistical Disclosure Control*, New York: Wiley. [203]
- Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014), "A Scalable Bootstrap for Massive Data," *Journal of the Royal Statistical Society, Series B*, 76, 795–816. [202]
- Leaver, V. (2009), "Implementing a Method for Automatically Protecting User-Defined Census Tables," *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*. [206]
- Lin, N., and Xi, R. (2011), "Aggregated Estimating Equation Estimation," *Statistics and Its Interface*, 4, 73–83. [204]
- Lubell-Doughtie, P., and Sondag, J. (2013), "Practical Distributed Classification Using the Alternating Direction Method of Multipliers Algorithm," in *2013 IEEE International Conference on 'Big Data'*, pp. 773–776. [202]
- McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, Boca Raton, FL: CRC Press. [xxxx]

- Minsker, S., Srivastava, S., Lin, L., and Dunson, D. (2014), "Scalable and Robust Bayesian Inference via the Median Posterior," in *Proceedings of the 31th International Conference on Machine Learning (ICML-14)*, pp. 1656–1664. [202]
- Miroshnikov, A., Wei, Z., and Conlon, E. M. (2015), "Parallel Markov Chain Monte Carlo for Non-Gaussian Posterior Distributions," *Stat*, 4, 304–319. [202]
- Narayanan, A., and Shmatikov, V. (2008), "Robust De-Anonymization of Large Sparse Datasets," in *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 111–125. [203]
- Neiswanger, W., Wang, C., and Xing, E. (2014), "Asymptotically Exact, Embarrassingly Parallel MCMC," in *Proceedings of the 30th International Conference on Uncertainty in Artificial Intelligence (UAI-14)*, pp. 623–632. [202]
- O'Keefe, C. M. (2008), "Privacy and the Use of Health Data-Reducing Disclosure Risk," *Electronic Journal of Health Informatics*, 3. [206]
- Perry, P. O. (2017), "Fast Moment-Based Estimation for Hierarchical Models," *Journal of the Royal Statistical Society, Series B*, 79, 267–291. [202]
- R Core Team (2014), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at <http://www.R-project.org/> [202,205,206]
- Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016), "Bayes and Big Data: The Consensus Monte Carlo Algorithm," *International Journal of Management Science and Engineering Management*, 11, 78–88. [202]
- Sweeney, L. (2002), "K-anonymity: A Model for Protecting Privacy," *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 10, 557–570. [203]
- White, T. (2009), *Hadoop: The Definitive Guide*, Sebastopol, CA: O'Reilly Media. [202]
- Xi, R., Lin, N., and Chen, Y. (2009), "Compression and Aggregation for Logistic Regression Analysis in Data Cubes," *IEEE Transactions on Knowledge and Data Engineering*, 21, 479–492. [204]
- Xu, M., Lakshminarayanan, B., Teh, Y. W., Zhu, J., and Zhang, B. (2014), "Distributed Bayesian Posterior Sampling via Moment Sharing," in *Advances in Neural Information Processing Systems*, pp. 3356–3364. [202]
- Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S., and Stoica, I. (2012), "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing," in *9th USENIX Symposium on Networked Systems Design and Implementation*, p. 2. [202]
- Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010), "Spark: Cluster Computing with Working Sets," *2nd USENIX Workshop on Hot Topics in Cloud Computing*. [202]