Disclosure Limitation in Longitudinal Linked Data

John M. Abowd
Cornell University, U.S. Census Bureau, CREST, and NBER

Simon D. Woodcock
Cornell University

August 20, 2001

Acknowledgement

The research reported in this paper was partially sponsored by the U.S. Census Bureau, the National Science Foundation (SES-9978093), and the French Institut National de la Statistique et des Etudes Economiques (INSEE) in association with the Cornell Restricted Access Data Center. The views expressed in the paper are those of the authors and not of any of the sponsoring agencies. The data used in this paper are confidential but the authors' access is not exclusive. No public use data sets were released as a part of this research. Restricted access to the French data was provided to Abowd by INSEE through an agreement with Cornell University. The authors thank Benoit Dostie, Sam Hawala, Janet Heslop, Paul Massell, Carol Murphree, Philip Steel, Lars Vilhuber, Marty Wells, Bill Winkler, and Laura Zayatz for helpful comments on earlier versions of this research.

1 Introduction

We consider longitudinal linked data, defined as microdata that contain observations from two or more related sampling frames with measurements for multiple time periods from all units of observation. Our prototypical longitudinal linked data set contains observations from work histories, and data on the individuals and employers observed in those work histories. We are primarily interested in the problem of confidentiality protection when data from all three sampling frames are combined for statistical analysis. Our goal is to develop and illustrate techniques that are appropriate for a variety of statistical analyses that are in widespread use in government agencies, like INSEE and the U.S. Census Bureau, and in academic research in the social sciences.

Current measures for confidentiality protection in linked data sets pose a number of problems for analysts. In particular, since the data sets that are linked are frequently constructed by different statistical agencies, the set of disclosure limitation requirements for the linked data are generally the union of disclosure limitation requirements of the several agencies. In practice, this can severely limit the usefulness of the linked data. These limitations on the usefulness of the resulting data motivate a unified approach to confidentiality protection in linked data.

In analyses of longitudinal linked data, analysts generally choose one of the underlying sampling frames as the reference population for the statistical modeling. Thus, the data matrix consists of rows that have been sampled from a specific population (*e.g.*, individuals, jobs, or employers) and the columns consist of functions of the linked data appropriate for that analysis (*e.g.*, sales/worker or the identity of the employing firm in an analysis of individuals; characteristics of the distribution of employees at several points in time for an analysis of employers). Confidentiality of any of the contributing data files can, thus, be compromised by elements

of either the rows or columns of the resulting data matrix. For example, linked individual information such as birth date or education can compromise the confidentiality of the individual data; linked information from the work history such as wage rates can compromise the confidentiality of both the individual and employer; linked information from the employer such as annual sales can compromise the confidentiality of the employer data. Our goal is to study methods that mask data from each of the source files in a manner that statistically preserves as much of the complicated relationships among the variables as possible.

1.1 Statistical concepts

We assume that the analyst is interested in the results of a statistical analysis of the form:

$$Y = f(X, \beta, \varepsilon) \tag{1}$$

where $\left[\begin{array}{cc} Y & X \end{array}\right]$ is the matrix of all available data (confidential and disclosable), $f(\cdot)$, is a (possibly) nonlinear function of X; β is a set of statistical parameters; and ε is a statistical error term distributed according to $p_{Y|X}(\varepsilon|X, \Omega)$. For completeness, note that X follows the joint distribution function $p_X(x|\Theta)$. We will consider methods for protecting the confidential data matrix $\left[\begin{array}{cc} Y & X \end{array}\right]$, primarily the use of multivariate multiple imputation techniques where $\left[\begin{array}{cc} Y & X \end{array}\right]$ is drawn from the predictive density, based on $p_{Y|X}(\varepsilon|X, \Omega)p_X(x|\Theta)$ and appropriate prior distributions on $\left(\begin{array}{cc} \Omega & \Theta \end{array}\right)$, to protect confidentiality for an entire analysis.

1.2 Background

Most statistical agencies assert that preserving confidentiality in longitudinal linked data and creating a statistically useful longitudinal public use product are incompatible goals.[1]. Perhaps for this reason, other researchers have not addressed the issue of disclosure limitation in longitudinal linked data. This is evident from the material presented in Appendix A, which contains a comprehensive, annotated bibliography of recent research on disclosure limitation. However, a number of authors have proposed methods for disclosure limitation for general microdata, some of which are directly relevant to our proposed method. Since these are discussed in detail in Appendix A, we only briefly summarize these works here.

(Kim and Winkler 1997) describe a two-stage masking technique applied to matched CPS-IRS microdata. The first stage of their technique is to mask variables with additive noise from a multivariate normal distribution with mean zero and the same correlation structure as the unmasked data. In the second stage, the authors randomly swap quantitative data within collapsed (age $\times$ race $\times$ sex) cells for records which pose an unacceptable disclosure risk. This approach preserves means and correlations in the subdomains on which the swap was done, and in unions of these subdomains. However, the swapping algorithm may severely distort means and correlations on arbitrary subdomains. Subsequent analysis of the masked data (e.g., (Moore 1996a) and (Winkler 1998)) indicates that the (Kim and Winkler 1997) approach adequately preserves confidentiality, and generates data which yield valid results for some analyses.

Our proposed approach draws heavily on the related suggestions of (Rubin 1993), (Fienberg 1994), and (Fienberg, Makov, and Steele 1998). These authors suggest releasing multiple data sets consisting of synthetic data; (Rubin 1993) suggests generating these data using multiple imputation techniques similar to those applied to missing data problems; (Fienberg 1994) suggests generating these data by bootstrap methods. There are numerous advantages to masking data via such methods. For example, valid statistical analyses of microdata masked by other methods generally require "not only knowledge of which masking techniques were used, but also special-purpose statistical software tuned to those masking techniques" (Rubin 1993, p. 461). In contrast, analysis of multiply-imputed synthetic data can be validly undertaken using

[1] See, *e.g.*, (Nadeau, Gagnon, and Latouche 1999) for a discussion of issues surrounding the creation of public use files for Statistics Canada's longitudinal linked Survey of Labour and Income Dynamics

standard statistical software simply by repeated application of complete-data methods. Furthermore, an estimate of the degree to which the disclosure proofing technique influences estimated model parameters can be inferred from between-imputation variability. Finally, since the released data are synthetic, *i.e.*, contain no data on actual units, they pose no disclosure risk. (Fienberg, Makov, and Steele 1998) have presented an application of such methods to categorical data; (Fienberg and Makov 1998) apply these ideas to develop a measure of disclosure risk.

In a series of related articles, (Kennickell 1991, 1997, 1998, 2000) describes the FRITZ algorithm, which is based on Rubin's (1993) suggestion and has been applied to disclosure limitation in cross-sectional survey data (the Survey of Consumer Finances, SCF). FRITZ is a sequential, iterative algorithm for imputing missing data and masking confidential data using a sequence of regression models (see Appendix A.3.2 for more details). In line with the above-mentioned proposals, the algorithm generates multiply-imputed, masked data. Unlike the suggestions of (Rubin 1993) and (Fienberg 1994), the released data are not synthetic. Rather, only a subset of cases and variables are masked, and the remaining data are left unmasked. The FRITZ algorithm has proven quite successful in application to the SCF, and for this reason we suggest its extension to longitudinal linked data.

1.3 Organization of the Paper

The organization of the paper is as follows. Section 2 presents the details of data masking and data simulation techniques applied to longitudinal linked data files. Section 3 summarizes the use of conventional complete-data methods for analyzing multiply-masked or simulated data. Section 4 applies our methods to confidential longitudinal linked data from the French national statistical institute. Section 5 provides a brief summary and conclusions. We include an extensive appendix that relates our methods to those already in the disclosure limitation literature.

2 Masking Confidential Data by Multiple Imputation

Consider a database with confidential elements Y and disclosable elements X. Both Y and X may contain missing data. Borrowing notation from (Rubin 1987), let the subscript mis denote missing data and the subscript obs denote observed data, so that $Y = (Y_{mis}, Y_{obs})$ and $X = (X_{mis}, X_{obs})$. We assume throughout that the missing data mechanism is ignorable.

The database in question is represented by the joint density $p(Y, X, \theta)$, where θ are unknown parameters. Following the related suggestions of (Rubin 1993) and (Fienberg 1994), the basic idea behind our disclosure limitation method is to draw masked data $\tilde{Y}$ from the posterior predictive density

$$p(\tilde{Y}|Y_{obs}, X_{obs}) = \int p(\tilde{Y}|X_{obs}, \theta) p(\theta|Y_{obs}, X_{obs})\, d\theta \tag{2}$$

to produce M multiply-imputed masked data files $\left(\tilde{Y}^m, X^m\right)$, .where $m = 1, ..., M$ In practice, it is simpler to first complete the missing data using standard multiple-imputation methods and then generate the masked data as draws from the posterior predictive distribution of the confidential data given the completed data. For example, first generate M imputations of the missing data (Y^m_{mis}, X^m_{mis}), where each implicate m is a draw from the posterior predictive density

$$p(Y_{mis}, X_{mis}|Y_{obs}, X_{obs}) = \int p(Y_{mis}, X_{mis}|Y_{obs}, X_{obs}, \theta)\, p(\theta|Y_{obs}, X_{obs})\, d\theta. \tag{3}$$

With completed data $Y^m = (Y^m_{mis}, Y_{obs})$ and $X^m = (X^m_{mis}, X_{obs})$ in hand, draw the masked data implicate

$\tilde{Y}^m$ from the predictive density

$$p(\tilde{Y}|Y^m, X^m) = \int p(\tilde{Y}|X^m, \theta) p\left(\theta|Y^m, X^m\right) d\theta \tag{4}$$

for each imputation m.

The longitudinal linked databases that we consider in this paper are very large and contain a variety of continuous and discrete variables. Furthermore, they are characterized by complex dynamic relationships between confidential and disclosable elements. For these reasons, specifying the joint probability distribution of all data, as in (3) and (4), is unrealistic. Instead, we approximate these joint densities using a sequence of conditional densities defined by generalized linear models. Doing so provides a simple way to model the complex interdependencies between variables that is computationally and analytically tractable. This method also provides a simple means of accommodating both continuous and categorical data by choice of an appropriate generalized linear model. We impute missing data in an iterative fashion using a generalization of Sequential Regression Multivariate Imputation (SRMI) developed by (Raghunathan, Lepkowski, Hoewyk, and Solenberger 1998). The SRMI approach and its generalization to the case of longitudinal linked data are described in the Section 2.1. Given the multiply-imputed completed data, we produce masked data on a variable-by-variable basis as draws from the posterior predictive distribution defined by an appropriate generalized linear model under an uninformative prior. Hence, if we let y_k denote a single variable among the confidential elements of our database, masked values $\tilde{y}_k$ are draws from

$$p\left(\tilde{y}_k|Y^m, X^m\right) = \int p(\tilde{y}_k|Y^m_{\sim k}, X^m, \theta) p\left(\theta|Y^m, X^m\right) d\theta \tag{5}$$

where $Y^m_{\sim k}$ are completed data on confidential variables other than y_k.

2.1 The SRMI Approach to Missing Data Imputation

For simplicity, consider a simple data set consisting of N observations on $K + P$ variables, ignoring for the moment the potential complications of longitudinal linked data. Let X be an $N \times P$ design or predictor matrix of variables with no missing values. Let Y be an $N \times K$ matrix of variables with missing values, and denote a particular variable in Y by y_k. Without loss of generality, assume they are ordered by their number of missing values, so that y_1 has fewer missing values than y_2, and so on, though the missing data pattern need not be monotone. Model-based imputations can use the density for Y given by

$$p\left(y_1, y_2, ..., y_K|X, \theta_1, \theta_2, ..., \theta_K\right) = p_1\left(y_1|X, \theta_1\right) p_2\left(y_2|X, y_1, \theta_2\right) ... p_K\left(y_K|X, y_1, y_2, ..., y_{K-1}, \theta_K\right) \tag{6}$$

where p_k are conditional densities and θ_k is a vector of parameters in the conditional density of y_k, $k = 1, ..., K$. The SRMI approach is to model each of these conditional densities using an appropriate generalized linear model with unknown parameters θ_k, then impute missing values by drawing from the corresponding predictive density of the missing data given the observed data. Again for simplicity, assume a diffuse prior on the parameters, *i.e.*, $\pi(\theta) \propto 1$.

The SRMI imputation procedure consists of L rounds. Denote the completed data in round $\ell + 1$ on some variable y_k by $y_k^{(\ell+1)}$. In round $\ell + 1$, missing values of y_k are drawn from the predictive density corresponding to the conditional density:

$$f_k\left(y_k|y_1^{(\ell+1)}, y_2^{(\ell+1)}, ..., y_{k-1}^{(\ell+1)}, y_{k+1}^{(\ell)}, ..., y_k^{(\ell)}, X, \theta_k\right) \tag{7}$$

where the conditional density f_k is specified by an appropriate generalized linear model, and θ_k are the parameters of that model. Hence under SRMI, at each round ℓ, the variable under imputation is regressed

on all non-missing data and the most recently imputed values of missing data. The imputation procedure stops after a predetermined number of rounds or when the imputed values are stable. Repeating the procedure M times yields M multiply-imputed data sets.

Note that if the missing data pattern is monotone (see (Rubin 1987)), then the imputations obtained in round 1 are approximate draws from the joint posterior predictive density of the missing data given the observed data. Furthermore, in certain cases the SRMI approach is equivalent to drawing from the posterior predictive distribution under a fully parametric model. For example, if all elements of Y are continuous and each conditional regression model is a normal linear regression with constant variance, then the SRMI algorithm converges to the joint posterior predictive distribution under a multivariate normal distribution with an improper prior for the mean and covariance matrix ((Raghunathan, Lepkowski, Hoewyk, and Solenberger 1998), p.11).

The SRMI method can be considered an approximation to Gibbs sampling. A Gibbs sampling approach to estimating (6) proceeds as follows. Conditional on the values $\theta_2^{(\ell)}, ..., \theta_K^{(\ell)}$ and $Y_1^{(\ell)}, ..., Y_K^{(\ell)}$ drawn in round ℓ, draw $\theta_1^{(\ell+1)}$ from its conditional posterior density, which is based on (6). Next, draw the missing values of y_1 conditional on the new value $\theta_1^{(\ell+1)}$, the completed data $X, y_2^{(\ell)}, ..., y_K^{(\ell+1)}$, and round ℓ parameter estimates $\theta_2^{(\ell)}, ..., \theta_K^{(\ell)}$. That is, in round $\ell + 1$ the missing values in y_k are drawn from:

$$p_k^* \left(y_k | X, y_1^{(\ell+1)}, ..., y_{k-1}^{(\ell+1)}, y_{k+1}^{(\ell)}, ..., y_K^{(\ell)}, \theta_1^{(\ell+1)}, ..., \theta_k^{(\ell+1)}, \theta_{k+1}^{(\ell)}, ..., \theta_K^{(\ell)} \right), \tag{8}$$

which is computed based on (6). Though such an approach is conceptually feasible, it is often difficult to implement in practice, especially when Y consists of a mix of continuous and discrete variables. SRMI approximates the Gibbs sampler to the extent that (7) approximates (8).

2.2 A Prototypical Longitudinal Linked Data Set

Before discussing the details of imputing missing values and masking longitudinal linked data, we must first introduce some basic notation. The prototypical longitudinal linked data set that we consider contains observations about individuals and their employers linked by means of a work history that contains information on the jobs each individual held with each employer. The data are longitudinal because complete work history records exist for each individual during the sample period and because longitudinal data exist for the employer over the same period. Suppose we have linked data on I workers and J firms with the following file structure. There are three data files. The first file contains data on workers, U, with elements denoted u_i, $i = 1, ..., I$. In the application below these data are time-invariant but in other applications they need not be. We refer to U as the individual characteristics. The second data file contains longitudinal data on firms, Z, with elements z_{jt}, $j = 1, ..., J$ and $t = 1, ..., T_j$. We refer to Z as the employer characteristics. The third data file contains work histories, W, with elements w_{it}, $i = 1, ..., I$ and $t = 1, ..., T_i$. The data U and W are linked by a person identifier. The data Z and W are linked by a firm identifier; we conceptualize this by the link function $j = J(i, t)$ which indicates the firm j at which worker i was employed at date t. For clarity of exposition, we assume throughout that all work histories in W can can be linked to individuals in U and firms in Z and that the employer link $J(i, t)$ is unique for each (i, t).[2]

2.3 Applying SRMI to Missing Data Imputation in Longitudinal Linked Data

With notation in hand, we now discuss applying SRMI to longitudinal linked data. The methods described in this Section and the next are applied to a particular linked longitudinal database in Section 4.

[2] The notation to indicate a one-to-one relation between work histories and indviduals when there are multiple employers is cumbersome. Our application properly handles the case of multiple employers for a given individual during a particular sample period.

When imputing missing data in each of the three files, we should condition the imputation on as much available information as possible. For example, when imputing missing data in the worker file U we should condition not only on the non-missing data in U (individual characteristics) but also on characteristics of the jobs held by the individual (data in W) and the firms at which the individual was employed (data in Z). Similarly, when conditioning the imputation of missing data in W and Z, we should condition on non-missing data from all three files. This necessitates some data reduction. To understand the data reduction, consider imputing missing data in the individual characteristics file U. Since individuals have work histories with different dynamic configurations of employers, explicitly conditioning the missing data imputation of individual characteristics on every variable corresponding to each job held by each worker is impractical–there are a different number of such variables for each observation to be imputed. A sensible alternative is to condition on some function of the available data which is well defined for each observation. For example, one could compute the person-specific means of time-varying work history and firm variables and condition the missing data imputation of variables in U on these.[3] Similar functions of person- and job-specific variables can be used to condition missing data imputation in the firm file Z. In what follows, we use the functions g, h, m and n to represent these data reductions.

It is also appropriate to condition the imputation of time-varying variables not only on contemporaneous data, but leads and lags of available data (including the variable under imputation). Because the dynamic configuration of work histories varies from worker to worker and the pattern of firm "births" and "deaths" varies from firm to firm, not every observation with missing data will have the same number of leads and lags available as conditioning variables. In some cases, there will be no leads and lags available at all. We suggest grouping observations by the availability of dynamic conditioning data (*i.e.*, the number of leads and lags available to condition missing data imputations) and separately imputing missing data for each group. This maximizes the set of conditioning variables used to impute each missing value. Again, some data reduction is generally necessary to keep the number of groups reasonable. For example, one might only condition on a maximum of s leads and lags, with $s = 1$ or $s = 2$. We parameterize the set of dynamic conditioning data available for a particular observation by κ_{it} in the work history file, and γ_{jt} in the firm file.

It may also be desirable to split the observations into separate groups on the basis of some observable characteristics, for example gender, full-time/part-time employment status, or industry. We parameterize these groups by λ_i in the individual file, μ_{it} in the work history file, and ν_{jt} in the firm file.

Given an appropriate set of conditioning data, applying SRMI to missing data imputation in longitudinal linked data is straightforward. The key aspects of the algorithm remain unchanged–one proceeds sequentially and iteratively through variables with missing data from all three files, at each stage imputing missing data conditional on all non-missing data and the most recently imputed values of missing data. As in the general case, the optimal imputation sequence is in increasing degree of missingness. As each variable in the sequence comes up for imputation, observations are split into groups based on the value of κ_{it}, γ_{jt}, λ_i, μ_{it}, and/or ν_{jt}. The imputes are drawn from a separate predictive density for each group. After the imputes are drawn, the source file for the variable under imputation is reassembled from each of the group files. Before proceeding to the next variable, all three files must be updated with the most recent imputations, since the next variable to be imputed may reside in another file (U, W, or Z). At the same time, the functions of conditioning data (including leads and lags) described above generally need to be re-computed. As in the general case, the procedure continues for a pre-specified number of rounds or until the imputed values are stable.

Explicitly specifying the posterior predictive densities from which the imputations are drawn is notationally cumbersome. For completeness, we give these in (9), (10), and (11). For a particular variable under imputation, subscripted by k, we denote by $U_{<k}$ the set of variables in U with less missing data than variable k; $W_{<k}$ and $Z_{<k}$ are defined analogously. We denote by $U_{>k}$ the set of variables in U with more missing

[3] Because the individual characteritics in our application are time-invariant, we use this approach but it is easy to generalize to the case where the individual characteristics (as distinct from the job characteristics) vary over time.

data than variable k, and define $W_{>k}$ and $Z_{>k}$ similarly. As in Section 2, we use the subscript *obs* to denote variables with no missing data. We also subscript conditioning variables by i, j, and t as appropriate to make clear the relationships between variables in the three data files. The predictive densities from which the round $\ell+1$ imputations are drawn are

$$\int f_{u_k}\left(\begin{array}{c} u_k|\, U^{(\ell+1)}_{<k,i}, U^{(\ell)}_{>k,i}, U_{obs,i}, g_k\left(\left\{Z^{(\ell+1)}_{<k,J(i,t)}, Z^{(\ell)}_{>k,J(i,t)}, Z_{obs,J(i,t)}\right\}_{t=1}^{t=T_i}\right), \\ h_k\left(\left\{W^{(\ell+1)}_{<k,it}, W^{(\ell)}_{>k,it}, W_{obs,it}\right\}_{t=1}^{t=T_i}\right), \lambda_i, \theta_k \end{array}\right) p_k\left(\theta_k|.\right) d\theta_k \tag{9}$$

$$\int f_{w_k}\left(\begin{array}{c} w_k|\, U^{(\ell+1)}_{<k,i}, U^{(\ell)}_{>k,i}, U_{obs,i}, \left\{Z^{(\ell+1)}_{<k,J(i,\tau)}, Z^{(\ell)}_{>k,J(i,\tau)}, Z_{obs,J(i,\tau)}\right\}_{\tau=t-s}^{\tau=t+s}, \\ \left\{w^{(\ell)}_{k,i\tau}\right\}_{\tau=t-s,\tau\neq t}^{\tau=t+s}, \left\{W^{(\ell+1)}_{<k,i\tau}, W^{(\ell)}_{>k,i\tau}, W_{obs,i\tau}\right\}_{\tau=t-s}^{\tau=t+s}, \kappa_{it}, \mu_{it}, \theta_k \end{array}\right) p_k\left(\theta_k|.\right) d\theta_k \tag{10}$$

$$\int f_{z_k}\left(\begin{array}{c} z_k|m_k\left(U^{(\ell+1)}_{<k,J^{-1}(i,t)}, U^{(\ell)}_{>k,J^{-1}(i,t)}, U_{obs,J^{-1}(i,t)}\right), \\ \left\{z^{(\ell)}_{k,j\tau}\right\}_{\tau=t-s,\tau\neq t}^{\tau=t+s}, \left\{Z^{(\ell+1)}_{<k,j\tau}, Z^{(\ell)}_{>k,j\tau}, Z_{obs,j\tau}\right\}_{\tau=t-s}^{\tau=t+s}, \\ n_k\left(\left\{W^{(\ell+1)}_{<k,J^{-1}(i,\tau)\tau}, W^{(\ell)}_{>k,J^{-1}(i,\tau)\tau}, W_{obs,J^{-1}(i,\tau)\tau}\right\}_{\tau=t-s}^{\tau=t+s}\right), \gamma_{jt}, \nu_{jt}, \theta_k \end{array}\right) p_k\left(\theta_k|.\right) d\theta_k, \tag{11}$$

where the posterior densities $p_k\left(\theta_k|.\right)$ are conditioned on the same information as the probability model for the k^{th} variable.

2.4 Masking the Completed Data

Repeating the missing data imputation method of the previous section M times yields M sets of completed data files (U^m, W^m, Z^m) which we shall call the completed data implicates $m = 1, ..., M$. The implicates are masked independently by drawing masked values of confidential data from an appropriate predictive distribution such as (5). We call the resulting M masked data files the masked data implicates $m = 1, ..., M$. Although in many ways the masking procedure is similar to the missing data imputation method described above, an important difference is that masking is not iterative. Masked data are drawn only once per observation-confidential variable-implicate triple.

As in the missing data imputation and for the same reasons, some data reduction is required when specifying the conditioning set for each confidential variable. Similarly, dynamic conditioning data (leads and lags) available for masking a particular variable will vary from observation to observation. Hence, as in the missing data imputation, it is useful to group observations by the set of such data available to condition the masking regressions. It is also useful to group observations on the basis of some key variables, such as gender, full time/part time employment status, and industry, for which we would expect parameters of the predictive distribution to differ. We retain the same notation for parameterizing these groups as defined above.

The masking algorithm for a single implicate is as follows. First, split each of the three files into groups as described above. Then, for each confidential variable, estimate an appropriate generalized linear model on each group, conditioning on a well chosen subset of the available data from all three files. Given the posterior distribution of the parameters of this generalized linear model, compute a draw from the predictive distribution for each confidential variable in each group. The masked data are these draws from the predictive distributions. The final step is to reassemble the masked data files from the various group files. Repeating this procedure on each completed data implicate yields multiply-imputed masked data. As before, the predictive densities are defined by an appropriate regression model and prior. For a given variable k from one of the source files we draw its masked implicate from the posterior predictive density corresponding

to an appropriate generalized linear model and an uninformative prior. For a particular implicate, these predictive densities are

$$\int f_{u_k}\left(u_k|\, U^m_{\sim k,i}, g_k\left(\left\{Z^m_{J(i,t)}\right\}_{t=1}^{t=T_i}\right), h_k\left(\{W^m_{it}\}_{t=1}^{t=T_i}\right), \lambda_i, \theta_k\right) p_k\left(\theta_k|.\right) d\theta_k \tag{12}$$

$$\int f_{w_k}\left(\begin{array}{c} w_k|\, U^m_i, \left\{Z^m_{J(i,\tau)}\right\}_{\tau=t-s}^{\tau=t+s}, \\ \left\{w^m_{k,i\tau}\right\}_{\tau=t-s,\tau\neq t}^{\tau=t+s}, \left\{W^m_{\sim k,i\tau}\right\}_{\tau=t-s}^{\tau=t+s}, \kappa_{it}, \mu_{it}, \theta_k \end{array}\right) p_k\left(\theta_k|.\right) d\theta_k \tag{13}$$

$$\int f_{z_k}\left(\begin{array}{c} z_k|\, m_k\left(\{U^m_i\}_{i\in\{i|j=J(i,t)\}}\right), \\ \left\{z^m_{k,j\tau}\right\}_{\tau=t-s,\tau\neq t}^{\tau=t+s}, \left\{Z^m_{\sim k,j\tau}\right\}_{\tau=t-s}^{\tau=t+s}, \\ n_k\left(\{W^m_{i\tau}\}_{\tau=t-s,i\in\{i|j=J(i,t)\}}^{\tau=t+s}\right), \gamma_{jt}, \nu_{jt}, \theta_k \end{array}\right) p_k\left(\theta_k|.\right) d\theta_k \tag{14}$$

where the posterior density of the parameters, $p_k\left(\theta_k|.\right)$, is conditioned on the same information as the conditional density of the variable being masked, and the subscript $\sim k$ refers to all other variables in the same source file. As always, there is a tradeoff between the analytic usefulness of the masked data file and the degree of confidentiality it affords. Below, we discuss various means of understanding the choices involved in these conflicting objectives.

2.4.1 Improving Confidentiality Protection

Our masking procedure preserves the configuration of the longitudinal histories in the three data files. That is, although all cases of confidential variables are masked, links between records in the three files are not perturbed. This preserves particular dynamic aspects of the database, such as individual work histories and firm births and deaths, as well as the history of worker-firm matches. In principle, the assumption of disclosable history configurations could be relaxed–for example, by perturbing some links between files, censoring some job records, or suppressing data on particular individuals or firms. We do not explore these issues in detail here, but note that perturbing the configuration of histories in the masked or completed data implicates may lead to substantial increases confidentiality protection.

A final step before releasing the masked data is to remove unique person and firm identifiers in the various data files. These can be replaced instead with randomly generated ones. Note that the identifiers used in the released data need not be the same in each implicate. In fact, using different identifiers in each implicate will serve to increase confidentiality protection, since this prevents an intruder from easily combining information about firms or individuals across implicates. To do so, records in each implicate would first need to be statistically matched to records from the other implicates.

2.4.2 Improving Analytic Usefulness

In most applications, substantial improvements in the analytic quality of the masked data can be achieved by imposing *a priori* restrictions on the masked values. In general, such restrictions will reduce between-implicate variability, and hence reduce the level of confidentiality protection. Restricting the masked values can be done in a variety of ways. Sampling from a posterior predictive density proceeds in two steps — first, sampling from the posterior density of model parameters and, second, sampling from the predictive density conditional on the parameter draw. Importance sampling of the parameters and/or masked values is one way to improve the analytic quality of the masked data. In cases where the data are highly collinear, the usual case in the type of data we are considering, estimates of parameter covariances are likely to be imprecise. In such cases, restricting parameter draws to central regions of the posterior density can

dramatically improve the quality of the masked data. Specifying parsimonious masking models will also be useful in such situations. We demonstrate an application of these methods in Section 4.

For a given parameter draw, restricting the draws of the masked values themselves will also serve to improve the analytic quality of the masked data. One approach is to restrict draws to central regions of the predictive density using standard methods. Another approach applicable to continuous variables is to individually restrict the masked values to lie inside an interval around true values. The interval can be specified in absolute or percentage terms. For example, one could restrict the masked values to lie within p percent of the completed values. To do so for a particular observation, sample from the predictive density until the restriction is satisfied, or until a pre-specified number of draws have been taken, at which time the masked value is set equal to one of the endpoints of the interval. An application of this method is described in Section 4.

Outliers provide two conflicting types of information. First, they may indicate that the original data have measurement errors (for example, earnings data that have been miscoded). Second, they may indicate that the underlying population is quite heterogeneous (for example, sales data within broadly defined industry groups). Either case has the potential to severely distort estimates of the masking equations and hence the masked data. We suggest treating outliers of the first type during the missing data imputation stage. Data values determined to be outliers of the first type can be set to missing and imputed along with other missing data. This procedure reduces the influence of these observations on both the missing data imputation and the data masking. It may substantially improve the analytic quality of the masked data. An important feature of our masking procedure, when combined with restricting the masked values to a range around the completed values, is that it is robust to outliers of the second type — the outlying values are perturbed in a manner consistent with the underlying data, without exerting undue influence on the masking of other observations.

We have not yet been explicit about the set of cases to be masked. In principle, not all observations need to be masked, though our method easily accommodates masking any number of cases in the three data files. Masking only a subset of cases will obviously improve the analytic usefulness of the data, though it does so at the expense of confidentiality protection. An example of such an application is (Kennickell 1997), who masks sensitive data on a small subset of cases in a cross-sectional file of individuals using methods related to those presented here. Details of the (Kennickell 1997) application can be found in the Appendix.

Traditional disclosure limitation methods may prove useful in preserving information when used in conjunction with our regression masking method. For example, some variables in the database may not pose a disclosure risk in aggregated analyses (*e.g.,* occupation or industry) but at a disaggregated level provide an easy means of compromising the confidentiality of records. For such variables, the overall analytic usefulness of the database may be better preserved by collapsing some cells or using other data coarsening methods than outright masking. We provide some examples below.

2.5 Simulated Data Based on Disclosable Summary Statistics

Disclosable summary statistics are defined as cross tabulations of discrete variables, conditional moments of continuous variables, generalized linear model coefficients, estimated covariance matrices, and estimated residual variances from such models. We construct disclosable summary statistics using an automated set of checks for conditions that are often associated with confidentiality preservation in tabulated data. Such checks normally include cell size and composition restrictions that generate primary suppressions as well as complementary suppressions generated by transformation tests that prevent the recovery of a suppressed cell from the released cells, conditional moments, or estimated model statistics. We build a data simulator that uses this disclosable statistical information to produce simulated draws from the predictive densities summarized by (12), (13), and (14). For comparability with our analyses of completed and masked data, we assume that there are also some variables X^m, possibly multiply-imputed for missing data, that can be

released in micro-data form. We note that if the variables in X^m are all discrete, then such a release is equivalent to releasing the full cross tabulation of all of the columns of X^m.

We provide M simulated draws from equation (4) using an approximation for the solution to (2). For each simulated implicate of Y we compute our approximation based on

$$p(\tilde{Y}|\ D\left(Y^m\right), X^m) = \int p(\tilde{Y}|\ D\left(Y^m\right), X^m, \theta) p\left(\theta|\ D\left(Y^m\right), X^m\right) d\theta \tag{15}$$

where the relation $D\left(Y^m\right)$ means that we have replaced the values of the original Y^m with aggregated values based on disclosable moments. To put the simulation procedure in context, we summarize the relation between the completed, masked and simulated data as follows. The original observed data are $\left(Y_{obs}, X_{obs}\right)$. The completed data are multiple imputations based on an approximation to (3). The masked data are multiple imputations conditional on the completed data and based on an approximation to equation (4). The simulated data are multiple imputations conditional on traditionally disclosable functions of the completed data and based on an approximation to equation (15).

The procedure we use to estimate (15) is analogous to the masking procedure described in Section 2.4. The data in each file U, W, and Z, are grouped according to the same conditions that are used to form $\lambda_i, \kappa_{it}, \mu_{it}$, and γ_{jt}, ν_{jt} in the data masking with the following exception. Each data configuration implied by these conditioning sets is subjected to an automatic traditional disclosure analysis that confirms that, within each file, the configurations are mutually exclusive, the cell sizes meet a minimum criterion, and there is no dominant unit. When cells fail such a test, they are collapsed. If no collapse is possible, the offending cell is suppressed. No marginal cells are used; hence, the margins constitute the complementary suppression where needed. To avoid notational clutter, we use the same symbols for these data configurations as in Section 2.4. For each data file and each data configuration within the file, we compute the conditional means of Y^m from completed data implicate m. We form $D\left(Y^m\right)$ by replacing, for each observation in each data file, the value of Y^m with the appropriate conditional mean.[4]

The exact simulation equations are given by

$$\int f_{u_k}\left(u_k|\ D\left(U^m_{\sim k,i}\right), D\left(g_k\left(\left\{Z^m_{J(i,t)}\right\}_{t=1}^{t=T_i}\right)\right), D\left(h_k\left(\left\{W^m_{it}\right\}_{t=1}^{t=T_i}\right)\right), \lambda_i, \theta_k\right) p_k\left(\theta_k|.\right) d\theta_k \tag{16}$$

which is based on equation (12),

$$\int f_{w_k}\left(\begin{array}{c} w_k|\ D\left(U^m_i\right), D\left(\left\{Z^m_{J(i,\tau)}\right\}_{\tau=t-s}^{\tau=t+s}\right), \\ D\left(\left\{w^m_{k,i\tau}\right\}_{\tau=t-s,\tau\neq t}^{\tau=t+s}, \left\{W^m_{\sim k,i\tau}\right\}_{\tau=t-s}^{\tau=t+s}\right), \kappa_{it}, \mu_{it}, \theta_k \end{array}\right) p_k\left(\theta_k|.\right) d\theta_k \tag{17}$$

which is based on equation (13),

$$\int f_{z_k}\left(\begin{array}{c} z_k|\ D\left(m_k\left(\left\{U^m_i\right\}_{i\in\{i|j=J(i,t)\}}\right)\right), \\ D\left(\left\{z^m_{k,j\tau}\right\}_{\tau=t-s,\tau\neq t}^{\tau=t+s}, \left\{Z^m_{\sim k,j\tau}\right\}_{\tau=t-s}^{\tau=t+s}\right), \\ D\left(n_k\left(\left\{W^m_{i\tau}\right\}_{\tau=t-s,i\in\{i|j=J(i,t)\}}^{\tau=t+s}\right)\right), \gamma_{jt}, \nu_{jt}, \theta_k \end{array}\right) p_k\left(\theta_k|.\right) d\theta_k, \tag{18}$$

which is based on equation (14), and where $p_k\left(\theta_k|.\right)$ is conditioned on $D\left(Y^m\right), X^m$. For each posterior distribution $p_k\left(\theta_k|.\right)$ we simulate a draw using M implicates based upon the generalized linear model statistics estimated for the appropriate masking equation. These statistics are also collapsed and suppressed to conform to the disclosure criteria used to form $D\left(Y^m\right)$.

[4] Additional moments can be used but we have not implemented this feature in our simulator.

3 Using the Completed, Masked and Simulated Data for Statistical Analysis

One of the principal advantages of multiply-imputed data is that valid statistical inferences can be obtained using standard complete-data methods. We illustrate these formulas for a generic statistic of interest, $\hat{Q}$, to be computed on multiple data implicates. The multiple implicates can be the result of missing data imputation (to produce completed data), masking (to produce masked data) or simulation (to produce simulated data). The formulas relating the complete-data methods and the multiple imputation methods, use standard relations derived in (Rubin 1987). For convenience, we reproduce these formulas here.

The quantity of interest, Q, may be either a scalar or a k-dimensional column vector. Assume that, with access to complete confidential data, inferences for Q would be based on

$$\left(Q - \hat{Q}\right) \sim N\left(0, V\right)$$

where $\hat{Q}$ is a statistic estimating Q, $N\left(0, V\right)$ is the normal distribution of appropriate dimension, and V the covariance of $\left(Q - \hat{Q}\right)$. Valid inferences can be obtained using the statistics $\hat{Q}$ and V computed on each of the data implicates. Denote the values obtained on each of the implicates by $\hat{Q}_1, ..., \hat{Q}_M$ and $V_1, ..., V_M$. The M complete-data statistics are combined as follows. Let

$$\bar{Q}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{Q}_m$$

denote the average of the complete-data estimates, and

$$\bar{V}_M = \frac{1}{M} \sum_{m=1}^{M} V_m$$

be the average of the complete-data variances. The between-implicate variance of the statistics $\hat{Q}_1, ..., \hat{Q}_M$ is

$$B_M = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{Q}_m - \bar{Q}_M\right) \left(\hat{Q}_m - \bar{Q}_M\right)^T$$

and the total variance of $\left(Q - \bar{Q}_M\right)$ is

$$T_M = \bar{V}_M + \frac{M+1}{M} B_M.$$

The standard error of a particular element of $\bar{Q}_M$ is the square root of the appropriate diagonal element of T_M. Examples of statistical analyses based on multiply-imputed masked data are given in the next section.

4 An Illustration Using French Longitudinal Linked Data

To illustrate the missing data imputation, masking, and simulation procedure described above, we apply these methods to a French longitudinal linked database on individuals and their employers. The data consist of both survey and administrative records collected by INSEE (Insitut National de la Statistique et des Etudes Economiques). The data structure is the same as the prototypical longitudinal linked data set described in Section 2.2. These data are described in detail in (Abowd, Kramarz, and Margolis 1999).

4.1 Individual and Work History Data

Individual characteristics and work history data are derived from the "Déclarations Annuelles des Données Sociales" (DADS), a large-scale administrative database of matched employer-employee information collected by INSEE. The work history data are based on mandatory employer reports of the gross earnings of each employee subject to French payroll taxes. These taxes apply to all "declared" employees and to all self-employed individuals — essentially all employed individuals in the economy.

The Division des Revenus prepares an extract of the DADS for scientific analysis which consists of all individuals employed in French enterprises who were born in October of even-numbered years, excluding civil servants. Our data span the years 1976 through 1996, with 1981, 1983, and 1990 excluded because the underlying administrative data were not collected in those years. Each record corresponds to a unique individual-year-establishment combination. Observations in the DADS file include an identifier which correspond to the employee (ID), an identifier that corresponds to the establishment (SIRET) and an identifier that corresponds to the economic enterprise of the establishment (SIREN). Since the employer data are reported at the enterprise level, we are concerned primarily with the enterprise identifier.

Since our purposes are mainly illustrative, we select a 20 percent random subsample of individuals in the DADS for the example application. A strict 10 percent subsample of the DADS, known as the Echantillon Démographique Permanent (EDP), includes detailed demographic information such as education. Our 20 percent subsample consists of the 10 percent of individuals in the EDP, plus an additional 10 percent random subsample of the other individuals in the DADS. The resulting subsample consists of 3,213,374 work history records on 362,913 individuals.

Time-invariant individual variables selected from the DADS for this illustration are gender, year of birth (range 1912 to 1980), and education. Time-varying job characteristics included are real annual compensation (annualized wage), occupation, geographic location of employment, full-time/other status, and number of days paid in the year (range 1 to 360).[5] Of these individual and work history variables, year of birth, education, real annual compensation, and days paid are selected for masking. Occupation is collapsed to five categories and geography is collapsed to two: employed in Ile-de-France (metropolitan Paris) and otherwise.

4.2 Firm Data

The primary source for our firm-level data is the "Enquête Annuelle d'Entreprises" (EAE) collected by INSEE and organized by SIREN. This survey collects detailed information from economically related establishments with the same owner (called enterprises) with annual year-end employment greater than 20. Variables selected from the EAE for this illustration are industry, annual sales, average employment over the year, and capital stock. Of these, annual sales, average employment, and capital stock are masked. Industry is collapsed to 40 categories prior to 1992, and 40 (different) categories thereafter.[6]

The sample of firms used for the example consists of firms in the EAE matched to work history records in our 20 percent subsample of the DADS. The firm sample is not representative of the French economy as a whole, which precludes certain analyses at the firm level. However, it serves to demonstrate the masking methods presented above. Our firm sample consists of 470,812 annual records on 105,813 enterprises.

We note that not all job records in the DADS can be linked to enterprises in the EAE. Non-matches arise when individuals are employed at enterprises with fewer than 20 employees and/or nonrespondent

[5] Days paid is an administrative variable that indicates the part of the year for which an employee received payments. Thus, by law, 360 days paid is a full-year work. Days paid from 1 to 359 represent payments for less than a full year of work. All individuals in the sample are permitted paid days of leave in accordance with French law and collective bargaining agreements, which cover more than 90 percent of all jobs.

[6] These categories correspond to standard French industrial classifications. In 1993, French industrial classifications changed from the Nomenclature d'Activités Productives (NAP) system to the Nomenclature d'Activités Francaises (NAF) system. There is no one-to-one mapping between these classification systems.

enterprises. This complicates the missing data imputation and masking since not all worker and job history records have firm data available as conditioning variables.

4.3 Missing Data Imputation Stage

Four of the variables selected for the illustration have missing data. These are education (missing for approximately half of individuals–those not in the EDP subsample), annual sales (missing in 47,796 matched EAE records), average employment (missing in 150,833 matched EAE records) and capital stock (missing in 35,989 matched EAE records). Hence the imputation sequence is capital stock, sales, employment, then education. This admits some computational efficiencies since the work history and person files only need to be updated with imputed firm data once per round (after the imputation of all three firm variables). Before estimating imputation models, observations are grouped as described in Section 2.3. For the firm variable imputations, these groups are defined purely by the availability of leads and lags (four groups).[7] For the wage outlier imputations, groups are defined by the availability firm data (due to non-matches), the availability of leads and lags, by gender, and by full-time/part-time status (20 groups). For the education imputations, we define groups on the basis of gender and the availability of firm data (four groups).

The firm variables with missing data are all continuous, so we use linear regression models for the imputation. Since all three are positive and highly skewed, these regression models are estimated in logarithms. Education is recorded in eight categories, so the appropriate imputation model is multinomial logistic regression.

The annualized wage variable was the only one with problematic outliers. Using the method described in Section 2.4.2, we set outlying values to missing and impute these along with other missing data. We detect outliers at the end of the first round of imputation on the other variables with missing data — the first point at which we have complete data upon which to condition an outlier detection model. Outliers are detected via a simple log wage regression. Wage values more than five standard deviations from their predicted value are considered outliers. After outliers are set to missing, the annualized wage variable joins the imputation sequence as the first variable imputed in round 2.

Initial experiments with imputing missing data in our database demonstrated the importance of specifying a parsimonious imputation model. For the logistic regressions, model selection is done manually. For the linear regressions we automate the model selection procedure. For each linear model estimated, we specify a set of candidate conditioning variables. The model is first estimated on all candidate variables. Only variables that meet the (Schwarz 1978) criterion are retained. The imputation model is then re-estimated on the reduced set of conditioning variables, and imputed values drawn from the corresponding predictive distribution. The set of candidate variables are selected along the lines described in Section 2.4. For the firm variables with missing data, candidate variables include up to one lead and lag of the variable under imputation (where available), contemporaneous values and up to one lead and lag of the other firm variables, firm-specific means of contemporaneous values of work history variables for employees, and mean characteristics of the workers employed at the firm in that period.[8] Candidate variables for imputing wage outliers include up to one lead and lag of the log annualized wage (where available), contemporaneous values and up to one lead and lag of other work history variables, contemporaneous firm variables for the firm at which the worker was employed, and worker characteristics. Conditioning variables for imputing missing education are manually selected from a candidate set of worker characteristics, and worker-specific means of work history and firm variables.

To further improve the quality of the imputed data when drawing from the predictive density, we restrict parameter draws for all estimated posterior distributions to lie within three standard deviations of the

[7] The number of groups given in this section correspond to rounds 2 through 10 of the imputation procedure. There are more groups in round 1 due to missing data on the variables that define the groups.

[8] For categorical variables in the work history and worker files, proportions in a given category are used in place of means.

posterior mode.[9] Initial experiments demonstrated remarkable improvements in the quality of the imputed data as a result of this restriction because it reduced collinearity of the conditioning data which increased the precision of posterior parameter covariances.

The imputation procedure consists of 10 rounds. Posterior parameter distributions in the imputation models change little after the sixth round. We repeat the procedure 10 times, yielding 10 completed data implicates.

4.4 Masking Stage

Confidential variables in each of the completed data implicates were masked using the methods described in Section 2.4. Observations were split into the same groups as described in the previous section. We used the same model selection techniques as in the missing data imputation, and restricted parameter draws in the same way. In addition, we restricted the masked values of continuous variables to lie within $p = 20$ percent of the true or imputed value. Masked values were re-drawn until they were within this interval; if after 100 draws the candidate implicate remained outside the interval, the masked value was set equal to the closest endpoint.

The models used to mask variables that had missing data were the same as those described above. The additional masked variables, year of birth and days paid, were both treated as continuous and masked using linear regression. The days paid variable takes values only in the interval between one and 360, so we apply a logit-like transformation to this variable for masking.[10] After masking, both year of birth and days paid were rounded to the nearest integer.

4.5 Simulation Stage

We simulated the same list of confidential variables as in the masking stage. The automatic disclosure proofing resulted in the suppression of data for 434 enterprise-years and the associated work histories. No individual data were suppressed. Observations were grouped according to the same methods used in the previous section. Parameter draws from the posterior distribution were restricted as in the imputation and masking stages. There is no access to the confidential micro data in the simulation stage, so the simulated values cannot be restricted to lie within an interval around the "true" value. Instead, the simulated values of days paid and year of birth were restricted to the observed sample range of these variables. The models used to simulate the variables were exactly the same models used to mask these variables, except that some models could not be used because they did not pass the disclosure tests.

4.6 Statistical Properties of the Completed, Masked and Simulated Data

Tables 1, 2 and 3 present basic univariate properties of confidential variables in the completed, masked and simulated data. Tables 1 and 2 present these for the individual and work history variables by gender. Table 3 presents statistics for the firm data. It is apparent that the masked and simulated data retain the basic univariate properties of the completed data. Biases in the means and variances of masked and simulated variables are generally small. In relative (percentage) terms, the bias is larger though still well with within acceptable limits. These biases are smaller in the masked data than the simulated data, and smaller for firm variables than individual and work history variables. The masking and simulation procedures lead to considerable relative increases in the variance of univariate statistics, as we would expect. These increases in variance are much more pronounced in the simulated data than in the masked data. In the masked firm

[9] For the logistic regressions, parameters are drawn from the normal approximation to the posterior density.

[10] This transformation is $logit(days\ paid) = \log\left(\frac{days\ paid}{365 - days\ paid}\right)$.

data, the variance of variable means and variances are at times lower than in the completed data. This is likely a result of our sampling frame for this illustration not being representative at the enterprise level.

Tables 4, 5 and 6 present basic bivariate properties of the confidential variables in the completed, masked and simulated data. The bivariate statistics are presented as tables of correlation coefficients (below the diagonal) and the between-implicate variance of the correlation coefficient (above the diagonal). These two statistics provide the most summary detail without cluttering the tables excessively. Table 4 presents the correlations among the time-invariant personal characteristics in the individual data file for both genders combined. Table 5 presents the correlations among the time-invariant and time-varying variables linked to the work history data for both genders combined. Table 6 presents the correlations for the firm-level data.

Tables 4 and 5 demonstrate that the masked data fully preserves the bivariate structure of the confidential data at the worker and work history levels. There is almost no bias and relatively little between-implicate variance complicating the inference about a correlation coefficient. The masked data does have substantially more between-implicate variance than the completed data; however, never enough to substantially affect inferences about the magnitude of the correlations. The simulated data preserves the correlation structure remarkably well given the limitations inherent in the simulation. There is more bias in the simulated data than in the masked data and the between-implicate variance is substantially greater than with the completed data but usually not large enough to affect the inference about the correlation coefficient.

Table 6 shows that the masked firm data substantially preserves the correlation structure with between-implicate variation comparable to the completed data. The simulated data display some biases (underestimation of the correlation coefficient) and substantially increased between-implicate variation. Given the sampling frame used to construct the firm-level data for this simulation, neither of these outcomes is surprising. The correlation structure of the simulated data is biased towards zero but not enough to substantially change economically meaningful conclusions about the bivariate relationships.

Given the structure of our imputation, masking and simulation equations, it is perhaps not surprising that the masked and simulated data preserve the first two moments so effectively. Our next analyses are based on models of substantive economic interest. The models predict variables in the work history file based on information in all three linked files. They thus provide a very stringent test of the scientific quality of the masked and imputed data for addressing questions about job-level outcomes.

4.6.1 Modeling Wages With Fixed Individual and Employer Effects

Our first substantive model predicts the log wage rate (real, full time, full year compensation) based on individual characteristics, employer characteristics, and unobservable individual and employer effects. We chose this example for two related reasons. First, this model can only be estimated using linked longitudinal employer-employee data (see (Abowd, Kramarz, and Margolis 1999)). Second, the dependent variable is among the most studied job-level outcomes in economics; hence, we can use substantial prior information to interpret the reasonableness of the estimated statistical models. We include only one observation per individual per time period, selecting the dominant job (based on days paid) for that time period.

The statistical model is a two-factor analysis of covariance with main effects only for the two factors. The covariates consist of time-varying characteristics of the individual, job, and employer. The first factor is an individual effect that is decomposed into a function of time-invariant personal characteristics and unobservable personal heterogeneity. The second factor is a firm effect that consists of unobservable employer heterogeneity. All components of the full design matrix of the model are non-orthogonal. We compute the full least squares estimator for all the effects by direct solution of the model normal equations using the conjugate gradient algorithm specialized to the sparse representation of our normal equations.[11] We

[11] Robert Creecy of the U.S. Census Bureau programmed the sparse conjugate gradient implementation as well as the graph-theoretic identification algorithm. Both programs are used by the Bureau's Longitudinal Employer-Household Dynamics program.

calculate the identifiable person and firm effects, for subsequent use in statistical analysis, using a graph-theoretic analysis of the group structure of the underlying design matrix (see (**?**)).

Table 7 presents coefficient estimates for the time-varying and time-invariant regressors in the model. All estimated regression coefficients are shown in the table. The completed data results are essentially the same as other analyses of these data (*e.g.* (Abowd, Kramarz, and Margolis 1999)). Wage rates increase with labor force experience at a decreasing rate. Profiles for women are flatter than those for men. There is a premium for working in Ile-de-France. Wage rates increase as the size of the firm increases (using log sales as the size measure). Often the quadratic term in the log sales relationship is negative but it is essentially zero in these data. As regards time-invariant personal characteristics, for men each schooling diploma increases earnings relative to the no-diploma reference group and the usual elementary school, middle school, high school, college/university, graduate school progression holds. Similar results hold for the women except that elementary school completion is less valuable than no-diploma.[12]

To consider the reliability of the estimates from the masked and simulated data, we first examine the experience profiles. Figure 1 shows the comparison of the completed, masked and simulated experience profiles for men and women. The horizontal axis is years of potential labor force experience (years since finishing school). The vertical axis is the log wage rate. The slope of the log wage profile, called the return to experience, is interpreted as the percentage change in the wage rate associated with a small increase in experience given the current level of experience. Although the masked coefficients differ from the completed data coefficients for both genders, the estimated profiles are essentially identical. The additional variation associated with the between-implicate component of the variance of the profile would not affect inferences about the experience profiles in any meaningful way. On the other hand, the profiles estimated from the simulated data are substantially flatter than those estimated from the completed or masked data and the profile for men is slightly flatter than the one for women. These are meaningful differences that would materially affect conclusions drawn from the simulated data. One might reasonably ask if there were indications in the simulated data analysis that the conclusions for this variable would be sensitive to the simulation. While the standard errors of the model coefficients are somewhat larger for the analysis of the simulated data, they are not enough larger to provide the necessary signal about the discrepancies shown in Figure 1.

The comparison of the log sales effect reveals that the masked data are once again quite close to the completed data and the standard errors of the masked data coefficients are larger than those of the completed data by a magnitude that allows the completed data estimate to fall within the usual posterior interval ranges of the masked data. The simulated data analysis of the log sales effect is substantially larger than the effect measured in either the completed or masked data. In contrast to the experience coefficients, however, there is plenty of evidence in the simulated data that the log sales effect has been unreliably estimated. Both the standard errors and the between-implicate component of variation indicate that this effect has been unreliably estimated in the simulated data.

Comparison of the estimated education effects for men and women reveal that the masked data yields reliable estimates of these effects. The simulated data yield acceptable results, with estimation uncertainty comparable to the masked data.

Table 8 compares correlations between the estimated effects in the completed, masked and simulated data. The correlations are computed over all observations in the work history file that enter the data analysis. The correlations in Table 8 are used to help understand the extent to which time-varying characteristics, time-invariant characteristics, unobservable person effects, and unobservable firm effects contribute to the explanation of log wage rates. Correlations between the estimated effects and log wages in the completed data indicate that person effects are somewhat more important than firm effects in explaining log wages. This is the usual result for such analyses using French data, and is accurately reproduced in both the masked

[12] In the EDP "no diploma" means that the respondent to the French census declared that he or she did not complete elementary school.

and simulated data. The person and firm effects are negatively correlated, again the usual result for French data and reliably reproduced by the masked and simulated data. All correlations with log wage rates are somewhat attenuated in the simulated data, where the estimated effects explain wage variation less well than in either the completed or masked data.

4.6.2 Modeling Full-Time/Part-Time Status

Our second substantive model predicts whether an employee in a particular job has full-time status based on characteristics of the individual (gender, labor force experience and education), the employer (sales, capital stock and employment), and the job (real wage rate, occupation, location of employment). This example was chosen for several reasons. First, the dependent variable is always observed and is among the non-confidential characteristics of the job in our masked and simulated data. Thus, the dependent variable has not been manipulated statistically in either the completed, masked or simulated data. Second, the model we use is a generalized linear model (logistic regression), which could be more sensitive to the linearity assumptions used to mask and simulate the continuous variables in the confidential data. Third, this variable is often not available in linked longitudinal employer-employee data and must be imputed. Thus, it is of substantive interest to estimate a statistical model for full-time status using a linked employer-employee data set in which the variable is measured.

Table 9 shows that the masked data does an excellent job of preserving inferences about the effects of individual, job and employer characteristics for predicting full-time status with very little increase in estimation error. The simulated data perform substantially less well; however, both the standard errors of the coefficients and the between-implicate component of variance signal the poorer performance of these data. The effects themselves are tricky to interpret because the equation is conditioned on the actual full-time wage rate in the job. Thus, the other effects must be interpreted as marginal effects, given the wage rate. For this reason we place more emphasis on the comparison across the completed, masked and simulated data sets — leaving the assessment of the reasonableness of any particular set of estimated effects to the reader.

5 Conclusions

Our goal was to provide a complete description of masking and data simulation algorithms that could be used to preserve the confidentiality of linked, longitudinal data while still providing data analysts with substantial information about the relationships in those data. We provided full implementation details for the application of these techniques to prototypical longitudinal linked employer-employee data. The data completion model is a full implementation of sequential regression multivariate imputation based on generalized linear models for all data with missing values. Our procedures generalize the existing methods to preserve the dynamic links among the individual, job, and employer characteristics. Our masking technique preserves confidentiality by replacing the confidential data with a draw from the predictive distribution of those data, given the values of the other confidential and non-confidential variables, where the predictive distribution exploits the modeling techniques used to complete the data. Finally, our simulation technique preserves confidentiality by replacing the confidential data with a draw from the predictive distribution of those data, given only disclosable summary statistics.

We apply our techniques to longitudinal linked data from the French national statistical institute. We show that our masking and simulating techniques do an excellent job of preserving first and second moments for the individual and work history variables. The performance on the firm-level data is not as good but this result is probably due to the way we specialized our methods to focus on the analysis of variables in the work history file. We believe that focusing our techniques on employer-level data, without insisting upon links to the work history or individual data would substantially improve their performance. The masked

data did an excellent job of reproducing the important statistical features of analyses of the wage rate and full-time employment status variables in the work history file. The simulated data did not perform as well as the masked data but did provide many useful statistical results. The simulated data results were reliable enough to be combined with restricted access to the confidential completed data as a part of a full research program.

A Appendix: Recent Research on Disclosure Limitation

In recent years, statistical agencies have seen an increasing demand for the data they collect, coupled with increasing concerns about confidentiality. This presents new challenges for statistical agencies, who must balance these concerns. Doing so requires techniques to allow dissemination of data that is both analytically useful and preserves the confidentiality of respondents.

This appendix presents recent research on disclosure limitation methods and concepts. Although our primary interest is in methods appropriate to longitudinal linked data, this topic has not been well-addressed in the literature. Hence, we review disclosure limitation methods and concepts appropriate to microdata in general. Since there are a number of reviews which provide a good summary of early research, (e.g., (Subcommittee on Disclosure Avoidance Techniques 1978a), (Subcommittee on Disclosure Avoidance Techniques 1994b), and (Jabine 1993)), we concentrate on recent research only.

The review is divided into four parts. The first presents general research on the disclosure limitation problem. The second presents research on measures of disclosure risk and harm. Part three discusses recent research into disclosure limitation methods for microdata, and the final part discusses the analysis of disclosure-proofed data.

A.1 General Research and Survey Articles

(Evans, Moore, and Zayatz 1996) This paper summarizes recent applications of a variety of disclosure limitation techniques to Census Bureau data, and outlines current research efforts to develop new techniques. The paper briefly discusses methods for microdata (including the two-stage additive noise and data-swapping technique of (Kim and Winkler 1997), the rank-based proximity swapping method of (Moore 1996b), and developing synthetic data based on log-linear models), methods under consideration for the 2000 Decennial Census, and methods for establishment tabular data (see the more detailed discussion of (Evans, Zayatz, and Slanta 1998) below).

(Fienberg 1997) This paper presents an excellent review of the disclosure limitation problem and recent research to address it. The paper is organized in eight sections. The first section is introductory. In the second, the author defines notions of confidentiality and disclosure, presents the debate between limited access versus limited data, and describes the role of the intruder in defining notions of disclosure and methods of disclosure limitation. The third section presents two detailed examples to illustrate the issues, namely issues surrounding the Decennial Census and the Welfare Reform Act. The fourth section classifies various disclosure limitation methodologies that have been proposed, and illustrates them in some detail. The fifth section considers notions of uniqueness in the sample and uniqueness in the population, and their role in defining notions of disclosure risk (see the discussion of (Fienberg and Makov 1998), (Boudreau 1995), and (Franconi 1999), below). The sixth section presents two integrated proposals for disclosure limitation: the ARGUS project (see the discussion of (Hundepool and Willenborg 1999) and (Nordholt 1999) below) and proposals for the release of simulated data (see Section A.3.2 below). The seventh section presents a brief discussion of issues pertaining to longitudinal data, and section 8 concludes.

(Winkler 1997) This paper briefly reviews modern record-linkage techniques, and describes their application in re-identification experiments. Such experiments can be used to determine the level of confidentiality protection afforded by disclosure limitation methods. The author stresses the power of such techniques to match records from disclosure-proofed data to other data sources. Emerging record-linkage techniques will allow re-identification in many existing public-use files, even though these files were produced by conscientious individuals who believed they were using effective disclosure limitation tools.

(National Research Council 2000) This book describes the proceedings of a workshop convened by the Committee on National Statistics (CNSTAT) to identify ways of advancing the often conflicting goals of exploiting the research potential of microdata and preserving confidentiality. The emphasis of the workshop was on linked longitudinal data – particularly longitudinal data linked to administrative data. The workshop addressed four key issues. These are: (1) the trade-off between increasing data access on the one hand, and improving data security on the other; (2) the ethical and legal requirements associated with data dissemination; (3) alternative approaches for limiting disclosure risks and facilitating data access – primarily the debate between restricting access and altering data; and (4) a review of current agency and organization practices. Some interesting observations in the report include:

- Researchers participating in the workshop indicated a preference for restricted access to unaltered data over broader access to altered data. However, these researchers also recognized the research costs associated with the former option.
- Although linking databases can generate new disclosure risks, it does not necessarily do so. In particular, many native databases are already sensitive, and hence require confidentiality protection. Linking may create a combined data set that increases disclosure risk, however a disclosure incident occurring from a linked data source is not necessarily caused by the linking. The breach of disclosure may have occurred in the native data as well.
- An appropriate measure of disclosure risk is a measure of the *marginal* risk. In other words, rather than comparing risks under various schemes with disclosure probability zero, one might consider the change in probability of disclosure as a result of a specific data release or linkage, or for adding or masking fields in a data set – the marginal risk associated with an action.
- In defense of data alteration methods, Fienberg noted that all data sets are approximations of the real data for a group of individuals. Samples are rarely representative of the group about which a researcher is attempting to draw inferences, rather it represents those for whom information is available. Even population data are imperfect, due to coding and keying errors, missing data, and the like. Hence, Fienberg finds the argument that perturbed data is not useful for intricate analysis not altogether compelling.

A.2 Measures of Disclosure Risk and Harm

(Lambert 1993) This paper considers various definitions of disclosure, disclosure risk, and disclosure harm. The author stresses that disclosure is in large part a matter of perception – specifically, what an intruder *believes* has been disclosed, even if it is false, is key. The result of a false disclosure may be just as harmful (if not worse) than the result of a true disclosure. Having distinguished between disclosure risk and disclosure harm, the author develops general measures of these.

The author defines two major types of disclosure. In an *identity disclosure* (or *identification*, or *re-identification*), a respondent is linked to a particular record in a released data file. Even if the intruder learns no sensitive information from the identification, it may nevertheless compromise the security of the data file, and damage the reputation of the releasing agency. To distinguish between true identification and an intruder's beliefs about identification, the author defines *perceived identification,* which occurs when an intruder believes a record has been correctly identified, whether or not this is the case. An *attribute disclosure* occurs when an intruder believes new information has been learned about the respondent. This may occur with or without identification. The *risk of disclosure* is defined as the risk of identification of a released record, and the *harm from disclosure* depends on what is learned from the identification.

Suppose the agency holds N records in a data file $\mathbf{Z}$, and releases a random sample $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ of n masked records on p variables. (Lambert 1993) defines several measures of perceived disclosure risk. A "pessimistic" risk of disclosure is given by:

$$D(\mathbf{X}) = \max_{1\leq j\leq N}\max_{1\leq i\leq n}\Pr\left[i^{th}\text{released record is } j^{th}\text{ respondent's record}|\mathbf{X}\right] \quad (19)$$
$$= \max_{1\leq j\leq N}\max_{1\leq i\leq n}\Pr\left[\mathbf{x}_i \text{ is } j^{th}\text{ respondent's record}|\mathbf{X}\right].$$

Minimizing the measure in (19) protects against an intruder looking for the easiest record to identify. Alternate measures of disclosure risk can also be defined on the basis of (19), for example:

$$D_{average}(\mathbf{X}) = \frac{1}{N}\sum_{j=1}^{N}\max_{1\leq i\leq n}\Pr\left[\mathbf{x}_i \text{ is } j^{th}\text{ respondent's record}|\mathbf{X}\right] \quad (20)$$

$$D_{total}(\mathbf{X}) = ND_{average}(\mathbf{X}) \ . \quad (21)$$

Equation (20) is a measure of data vulnerability based on the average risk of perceived disclosure, whereas (21) is a measure of the cumulative risk. An alternate measure of the total risk of perceived identification can be defined as the number of cases for which the risk of perceived disclosure exceeds a threshold τ :

$$D_{\tau}(\mathbf{X}) = \#\left\{j : \max_{1\leq i\leq n}\Pr\left[\mathbf{x}_i \text{ is } j^{th}\text{ respondent's record}|\mathbf{X}\right] \geq \tau\right\} .$$

The author proceeds to develop several detailed examples, and provide a general measure of disclosure harm, which is not presented here.

(Fienberg and Makov 1998) This paper reviews several concepts, namely uniqueness in sample, uniqueness in the population, and some notions of disclosure. The main contribution is a proposed approach for assessing disclosure potential as a result of sample uniqueness, based on log-linear models. A detailed description of this method follows.

Suppose a population is cross-classified by some set of categorical variables. If the cross-classification yields a cell with an entry of "1" then the individual associated with this entry is defined as a *population unique*. Population uniqueness poses a disclosure risk, since an intruder with matching data has the potential to match his or her records against those of the population unique. This creates the possibility of both re-identification and attribute disclosure.

A *sample unique* is defined similarly – an individual associated with a cell count of "1" in the cross-classification of the sample data. Population uniques are also sample uniques if they are selected into the sample, but being a sample unique does not necessarily imply being a population unique. The focus of the (Fienberg and Makov 1998) approach for assessing disclosure potential is to use uniqueness in the sample to determine the probability of uniqueness in the population. Note that sample uniqueness is not necessarily required for such an endeavor – "small" cell counts may also pose a disclosure risk. For example, a count of "2" may allow an individual with almost unique characteristics to identify the only other individual in the sample with those characteristics. If the intruder did not also possess these characteristics, then a cell count of "2" could allow the individuals to be linked to the intruder's data with probability 1/2. The extension to larger yet still "small" cell counts is obvious.

Let N denote the population size, n the size of the released sample, and K the maximum number of "types" of individuals in the data, as defined by the cross-classifying variables (*i.e.*, the total number of cells). Let F_i and f_i, $i = 1, ..., K$, denote the counts in the cells of the multiway table summarizing the entire population and sample, respectively. Then a crucial measure of the vulnerability of the data is given by:

$$\sum_{i=1}^{K} \Pr\left(F_i = 1 | \, f_i = 1\right) \tag{22}$$

Most prior attempts to estimate (22) assumed distributions for F_i and f_i (e.g., (Bethlehem, Keller, and Pannekoek 1990) and (Skinner and Holmes 1993)). The (Fienberg and Makov 1998) approach differs by assuming the released sample is drawn from a population with cell probabilities $\left\{\pi_i^{(N)}\right\}$ which follow a log linear model (including terms such as main effects interactions), of the form

$$\log(\pi_i^{(N)}) = g_N(\boldsymbol{\theta}_i) \tag{23}$$

where $\boldsymbol{\theta}_i$ are parameters. The authors propose to fit (23) to the observed counts $\{f_i\}$. Denote the estimated cell probabilities $\left\{\hat{\pi}_i^{(n)}\right\}$. Ideally, one would like to develop analytical formulae for $\Pr(F_i = 1 | f_i = 1)$, but this will frequently be infeasible since many of the log-linear models that result from the estimation process will not have a closed-form representation in terms of the minimal marginal sufficient statistics. Instead, the authors propose the following simulation approach. First, use the records on the n individuals in the sample $(x_1, ..., x_n)$ to generate $(N - n) \times H$ records from $\left\{\hat{\pi}_i^{(n)}\right\}$. This results in H populations of size N, each containing $(N - n)$ "new" records obtained by some form of imputation (*e.g.*, see (Little and Rubin 1987)), or multiple imputations from some posterior distribution (*e.g.*, see (Rubin 1987)). Next, let $\bar{F}_i(j) = \bar{F}_i(x_1, ..., x_N, j)$ be the count in cell i of the jth imputed population. Similarly, let $\bar{f}_i = \bar{f}_i(x_1, ..., x_N)$ be the count in cell i of the released data (the sample). Clearly, $\bar{f}_i \neq 1 \Longrightarrow \bar{F}_i(j) \neq 1$. We can estimate (22) by:

$$\sum_{i=1}^{K} \widehat{\Pr}(F_i = 1 | f_i = 1) = \sum_{i=1}^{K} \sum_{j=1}^{H} \frac{\mathbf{1}\left[\left(\bar{F}_i(j) = 1\right) \cap \left(\bar{f}_i = 1\right)\right]}{H} \tag{24}$$

where the function $\mathbf{1}[A] = 1$ if A is true, and zero otherwise. Equation (24) can be used to assess the disclosure risk of the released data for a given release size n. Since (24) is likely to decrease as $(N - n)$ increases, the statistical agency is motivated to reduce n to the point that (24) indicates disclosure is infeasible. Note that if we remove the summation over i in (24), then we can obtain a cell-specific measure of disclosure risk.

(Fienberg and Makov 1998) do not address the sample error of the estimate in (24). They also do not address the inherent trade-off that an agency faces when choosing n based on (24) between reduced disclosure risk and increased uncertainty in the released data.

(Boudreau 1995) This paper presents another measure of disclosure risk based on the probability of population uniqueness given sample uniqueness. For the case of microdata containing discrete key variables, the author determines the exact relationship between unique elements in the sample and those in the population. The author also gives an unbiased estimator of the number of population uniques, based on sample data. Since this estimator exhibits great sampling variability for small sampling fractions, the author models this relationship. After observing this conditional probability for a number of real populations, the author provides a parametric formulation of it. This formulation is empirical only – it has not theoretical justification. However, the empirical formulation is much more flexible than earlier measures of disclosure risk based on uniqueness which required distributional assumptions (e.g. the Poisson-Gamma model of (Bethlehem, Keller, and Pannekoek 1990) or the Poisson-Lognormal model of (Skinner and Holmes 1993)).

(Willenborg and Kardaun 1999) This paper presents an alternate measure of disclosure risk appropriate to microdata sets for research (as opposed to public use files). In such files there is generally no

requirement that all records be absolutely safe, since their use is usually covered by a contractual agreement which includes a non-record-matching obligation. The approach is to define a measure of the "degree of uniqueness" of an observation, called a *fingerprint*. A fingerprint is a combination of values of identifying (key) variables that are unique in the data set at hand, and contain no proper subset with this property (so it is a minimum set with the uniqueness property). The authors contend that records with "many" "short" fingerprints (i.e., fingerprints comprised of a small number of variables) are "risky", and should not be released. Appropriate definitions of "many," "short" and "risky" are at the discretion of the data collection/dissemination agency. In this way, defining disclosure risk in terms of fingerprints is very flexible. The authors propose that agencies use the fingerprinting criterion to identify risky records, and then apply disclosure-limitation measures to these records. The paper contains a discussion of some design criteria for an implementation of the fingerprinting criterion, and stipulates some useful heuristics for algorithm design.

(Franconi 1999) This paper reviews recent developments in measures and definitions of disclosure risk. The author stresses differences between methods appropriate to social data and to business data. These differences are due to differences in the underlying data. Namely, social data are generally from large populations, have an inherent dependent structure (*i.e.*, groups such as families or households exist in the data), and are characterized by key variables of a categorical nature. These characteristics allow one to tackle the disclosure limitation problem via concepts of uniqueness. Business data, on the other hand are generally from small populations, with a skewed distribution, and have key variables which are primarily continuous. Uniqueness concepts are generally not useful here, since nearly all cases would be considered unique. In both cases, the author stresses the need to take account of hierarchies in the data, such as the grouping of cases into families and households. These hierarchies provide additional information to an intruder attempting to identify records, hence they should be incorporated into measures of disclosure risk.

A.3 Disclosure Limitation Methods for Microdata

A.3.1 Additive Noise Methods

(Fuller 1993) This paper considers a variety of masking methods in which error is added to data elements prior to release. These fall generally within the class of measurement error methods. The author stresses that to obtain consistent estimates of higher-order moments of the masked data and functions of these moments such as regression coefficients, measurement error methods and specialized software are required. Other techniques, such as data switching and imputation, can produce biased estimates of some sample covariances and other higher-order moments. The approach is related to that of (Kim and Winkler 1997), but applicable to data which is not necessarily multivariate normal.

(Kim and Winkler 1997) This paper presents a two-stage disclosure limitation strategy, applied to matched CPS-IRS data. The disclosure concern in this data arises from the match: the CPS data are already masked, but the IRS tax data is not. The IRS data need to be sufficiently well-masked so they cannot easily be used in re-identifications, either alone or in conjunction with unmasked key variables from the CPS. The procedure is as follows.

The data in question are known to be approximately multivariate normal. Hence, in the first stage noise from a multivariate normal distribution with mean zero and the same correlation structure as the unmasked data is added to the IRS income variables. As discussed in (Little 1993) and (Fuller 1993), such an approach is currently the only method that preserves correlations. Following the addition of noise to the data, the authors determine the re-identification risk associated with the data by matching the raw linked data to the masked file. In cases where the re-identification risk was deemed too great, the authors randomly swap quantitative data within collapsed (age $\times$ race $\times$ sex) cells. This approach preserves means and correlations in the subdomains on which the swap was done, and in unions of these subdomains. However, the swapping

algorithm may severely distort means and correlations on arbitrary subdomains. Finally, the authors assess both the confidentiality protection offered by their method and the analytic usefulness of the resulting files, and conclude that both are good.

(Moore 1996a) This paper provides a critical examination of the degree of confidentiality protection and analytic usefulness provided by the (Kim and Winkler 1997) method. The author concludes that the method is both useful and highly feasible. The author also considers some particular aspects of the algorithm, such as optimal parameter values which generate "sufficient" masking with minimal distortion to second moments. Finally, the author considers how much masking is "sufficient," given reasonable assumptions on intruder knowledge, tools, and objectives.

(Winkler 1998) This paper compares the effectiveness of a number of competing disclosure limitation methodologies to preserve both confidentiality and analytic usefulness. The methods considered include the additive-noise and swapping techniques of (Kim and Winkler 1997), the additive-noise approach of (Fuller 1993), and μ-ARGUS suppression as described in (Hundepool and Willenborg 1999) and (Nordholt 1999) in Section A.3.3. The author arrives at several conclusions. First, the (Fuller 1993) additive-noise method may not provide as much protection as that author had originally suggested. In particular, sophisticated matching techniques may allow for a significantly higher re-identification rate than previously thought. Second, a naive application of μ-ARGUS to the linked CPS-IRS data described in (Kim and Winkler 1997) did little to preserve either confidentiality or analytic usefulness. More sophisticated methods, including a variant of the (Kim and Winkler 1997) method that included a μ-ARGUS pass on the masked data, were much more successful. The authors conclude that additive-noise methods can produce masked files that allow some analyses to approximately reproduce the results obtained with unmasked data. When additional masking procedures are applied such as limited swapping or probability adjustment ((Fuller 1993)), then disclosure risk is significantly reduced, though analytic properties are somewhat compromised.

(Duncan and Mukherjee 1998) This paper derives an optimal disclosure limitation strategy for statistical databases – *i.e.*, micro-databases which respond to queries with aggregate statistics. As in all disclosure limitation problems, the aim is to maximize legitimate data access while keeping disclosure risk below an acceptable level. The particular confidentiality breach considered is called a *tracker attack*: a well known intruder method in databases with query set size (QSR) control. QSR control is a query restriction technique where a query is disallowed if the number of records satisfying the query is too small (or too large, by inference from the complementary query). A tracker attack is a finite sequence of legitimate queries that yields the same information as a query precluded under QSR. The authors show that the optimal method for thwarting tracker attacks is a combination of query restriction and data masking based on additive noise. The authors also derive conditions under which autocorrelated noise is preferable to independent noise or "permanent" data perturbation.

(Evans, Zayatz, and Slanta 1998) This paper presents an additive-noise method for disclosure limitation which is appropriate to establishment tabular data. The authors propose adding noise to the underlying microdata prior to tabulation. Under their approach, "more sensitive" cells receive more noise than less sensitive cells. There is no attempt to preserve marginal totals. This proposal has numerous advantages over the cell-suppression approach which is usually applied to such data. In particular, it is far simpler and less time-consuming than cell-suppression techniques. It also eliminates the need to coordinate cell suppressions between tables, and eliminates the need for secondary suppressions, which can seriously reduce the amount of information in tabular releases. The authors also contend that an additive noise approach may offer more protection than cell-suppression, although suppression may give the appearance of offering more protection.

(Pursey 1999) This paper discusses the disclosure control methods developed and implemented by Statistics Canada to release a Public Use Microdata File (PUMF) of financial data from small businesses. This is a fairly unique enterprise – in most cases, statistical agencies deem it too difficult to release public use microdata on businesses that preserve confidentiality. The paper discusses the five steps taken to create the PUMF: (1) make assumptions about an intruder's motivation, information, and tools; (2) make disclosure control goals based on these assumptions; (3) translate these goals into mathematical rules; (4) implement these rules to create the PUMF; and (4) measure the data quality of the PUMF. These are discussed briefly below.

It is assumed that an intruder seeks to identify any record in the PUMF, and has access to the population data file from which the PUMF records are drawn. It is assumed that identification is achieved via nearest-neighbor matching to the population file. Given these assumptions, the following disclosure control goals were set:

- Ensure a low probability that a business from the population appears in the PUMF (less than $r\%$), and that an intruder cannot determine that a particular business is in the PUMF.
- Ensure that each continuous variable is perturbed and that an intruder cannot undo the perturbation.
- Ensure a low probability that a PUMF record can be correctly linked to itself in the population file (less than $p\%$), and that an intruder cannot determine whether a link has been correctly or incorrectly made.
- Remove unique records.

Continuous variables were perturbed according to methods similar to those of (Kim and Winkler 1997). First, independent random noise was added to each datum, subject to the constraints that the minimum and maximum proportion of random noise is constant for each datum, and that within a record the perturbations are either always positive or always negative. Next, the three highest data values of each variable in each cell were replaced with their average. Finally, all data values were rounded to the nearest \$1000. Since a less than $p\%$ linkage rate was deemed necessary, in industry cells with a correct linkage rate greater than $p\%$, the data was further perturbed by data swapping with the second-nearest neighbor until a $p\%$ linkage rate was achieved.

After implementing the above disclosure control methods, the resulting data quality was analyzed. The general measure used was one of relative distance: $Rd = (x_a - x_b)/(x_a + x_b)$, where x_a is data or a sample statistic after disclosure control, and x_b is the same data or sample statistic before disclosure control. All variables in the PUMF and a variety of sample statistics were analyzed according to this distance measure. The results indicated that the resulting data quality was good to fair for unincorporated businesses, fair to poor for incorporated businesses.

A.3.2 Multiple Imputation and Related Methods

(Rubin 1993) (Rubin 1993) is the first paper to suggest the use of multiple imputation techniques for disclosure limitation for microdata analyses. His radical suggestion – to release only synthetic data generated from actual data by multiple imputation – is motivated by the forces outlined at the outset of this review. Namely, an increase in the demand for public use microdata, and increasing concern about the confidentiality of such data.

Rubin's (1993) approach has a number of advantages over competing proposals for disclosure limitation, such as microdata masking. For example, valid statistical analyses of masked microdata generally require "not only knowledge of which masking techniques were used, but also special-purpose statistical software tuned to those masking techniques" ((Rubin 1993, p. 461)). In contrast, analysis of multiply-imputed

synthetic data can be validly undertaken using standard statistical software simply by using repeated applications of complete-data methods. Furthermore, an estimate of the degree to which the disclosure proofing techniques influence estimated model parameters can be obtained from between-imputation variability. Finally, since the released data is synthetic, *i.e.*, contains no data on actual units, it poses no disclosure risk.

The details of Rubin's (1993) proposal are as follows. Consider an actual microdata sample of size n drawn using design D from a much larger population of N units. Let X represent background variables (observed, in principle, for all N units), Z represent outcome variables with no confidentiality concerns, and Y represent outcome variables with some confidentiality concerns. Note that Z and Y are only observed for the n sampled units, and missing for the $N-n$ unsampled units. A multiply-imputed population consists of the actual X data for all N units, the actual $\left[\begin{array}{cc} Z & Y \end{array}\right]$ data for the n units in the sample, and M matrices of $\left[\begin{array}{cc} Z & Y \end{array}\right]$ data for the $N-n$ unsampled units, where M is the number of multiple imputations. The multiply-imputed values of $\left[\begin{array}{cc} Z & Y \end{array}\right]$ are obtained from some model with predictors X. Given such a multiply-imputed population and a new survey design D^* for the microdata to be released (possibly the same as D), the statistical agency can draw a sample of $n^* \ll N$ units from the multiply-imputed population which is structurally like an actual microdata sample of size n^* drawn from the actual population using design D^*. This can be done M times to create M replicates of the $\left[\begin{array}{cc} Z & Y \end{array}\right]$ values. To ensure that no actual data are released, the statistical agency could draw the samples from the multiply-imputed population excluding the n actual units.

(Rubin 1993) recognizes the information loss inherent in the multiple-imputation technique. However, some aspect of this information loss are subtle, and he presents these as the following two facts. First, although the actual $\left[\begin{array}{cc} Z & Y \end{array}\right]$ and the population values of X contain more information than the multiply-imputed population, if the imputation model is correct, then as M increases, the information in the latter is essentially the same as in the former. Second, the information in the original microdata sample of size n may be greater than, less than, or equal to the information in the multiply-imputed sample of size n^*; the relationship will depend on the estimand under investigation, the relative sizes of n and n^*, the magnitude of M, the designs D and D^*, and the ability of X to predict $\left[\begin{array}{cc} Z & Y \end{array}\right]$.

(Fienberg 1994) (Fienberg 1994) proposes a method of confidentiality protection in the spirit of (Rubin 1993). Whereas (Rubin 1993) suggests generating synthetic microdata sets by multiple imputation, (Fienberg 1994) suggests generating synthetic microdata by bootstrap methods. This method retains many of the desirable properties of Rubin's (1993) proposal – namely disclosure risk is reduced because only synthetic data are released, and the resultant microdata can be analyzed using standard statistical methods.

To discuss the details of his proposal, let us restate the statistical agency's problem. As before, suppose the agency collects data on a random sample of size n from a population of size N (ignore aspects of the sample design). Let F be the true p-dimensional c.d.f. of the data in the population, and let $\bar{F}$ be the empirical c.d.f. based on the sample of size n. The disclosure problem arises because researchers request, in essence, access to the full empirical p-dimensional c.d.f., $\bar{F}$. Because of guarantees of confidentiality, the agency believes it cannot release $\bar{F}$ since an intruder may be able to identify one or more individuals in the data.

Fienberg's (1994) proposal is as follows. Suppose the statistical agency has a "smoothed" estimate of the c.d.f., $\hat{F}$, derived from the original sample c.d.f. $\bar{F}$. Rather than releasing either $\bar{F}$ or $\hat{F}$, the agency could sample from $\hat{F}$ and generate a synthetic bootstrap-like sample of size n. Denote the empirical c.d.f. of the synthetic microdata file as $\bar{G}$. (Fienberg 1994) notes some technical details surrounding $\bar{G}$ which have yet to be addressed. Namely, under what conditions would replicates of $\bar{G}$, say $\bar{G}_i$ for $i = 1, ..., B$, be such that as $B \to \infty$, $\frac{1}{B}\sum_{i=1}^{B} \bar{G}_i \to \hat{F}$? Is a single replicate sufficient, or would multiple replicates be required for valid analyses, or possibly the average of multiple replicates? Bootstrap theory may provide some insight into these issues.

(Fienberg, Makov, and Steele 1998) The authors reiterate Feinberg's (1994) proposal for generating synthetic data via bootstrap methods, and present a related application to the case of categorical data. Categorical data can be represented by a contingency table, for which there is a direct relationship between a specific hierarchical loglinear model and a set of marginal tables that represent the minimal sufficient statistics of the model. The authors present an example of a three-way table, for which they obtain maximum likelihood estimates of the expected cell values under a loglinear model. The suggestion is to release the MLEs as a public use product, rather than the actual data. They then generate 1,000,000 tables with the same two-way margins, and perform a goodness-of-fit test based on the MLEs. They find that the sparseness of the table in their example presents some problems for accurate loglinear modeling.

In a comment to this article, (Kooiman 1998) expresses doubt as to the feasibility of the (Fienberg 1994) and (Fienberg, Makov, and Steele 1998) proposal for generating synthetic data. He makes a connection between the proposed method and a data-swapping exercise subject to fixed margins. (Kooiman 1998) shows that for large data sets with many categorical variables and many categories, such an exercise is likely impossible. He also finds the relationship between the synthetic data proposal and the categorical data example tenuous, at best.

(Kennickell 1991), (Kennickell 1997), (Kennickell 1998), (Kennickell 2000) In a series of articles, (Kennickell 1991), (Kennickell 1997), (Kennickell 1998), (Kennickell 2000), describes the Federal Reserve Imputation Technique Zeta (FRITZ), used for both missing value imputation and disclosure limitation in the Survey of Consumer Finances (SCF). The SCF is a triennial survey administered by the Federal Reserve Board to collect detailed information on all household assets and liabilities. Because holdings of many types of assets are highly concentrated in a relatively small fraction of the population, the SCF heavily oversamples wealthy households. Since such households are likely to be well-known, at least in their localities, the data collection process presents a considerable disclosure risk. As a first step towards implementing the proposal of (Rubin 1993), the SCF simulates data for a subset of sample cases, using the FRITZ multiple imputation algorithm. This approach is highly relevant for our current research, and hence we discuss it in some detail here.

Using The FRITZ Algorithm for Missing Data Imputation As mentioned above, the FRITZ algorithm is used both for missing value imputation and disclosure limitation in the SCF. The algorithm is most easily understood in the context of missing data imputation. We return to the issue of its application to disclosure limitation below.

The FRITZ model is sequential in the sense that it follows a predetermined path through the survey variables, imputing missing values one (occasionally two) at a time. The model is also iterative in that it proceeds by filling in all missing values in the survey data set, using that information as a basis for imputing the following round, and continuing the process until key estimates are stable. Five imputations are made for every missing value, hence the method is in the spirit of Rubin's (1993) proposal. The following describes the FRITZ technique for imputing missing continuous variables.

For convenience, suppose the iterative process has completed $\ell-1$ rounds, and we are currently somewhere in round ℓ, with a data structure as given below: (reproduced from (Kennickell 1998))

$$\left[\begin{array}{cccc} & \text{Iteration } \ell-1 & & \\ y_1 & \chi_{11}^{\ell-1} & x_{12} & x_{13} \\ \Psi_2^{\ell-1} & \chi_{21}^{\ell-1} & x_{22} & \chi_{23}^{\ell-1} \\ y_3 & x_{31} & x_{32} & x_{33} \\ \ldots & & & \\ y_{n-2} & x_{n-2,1} & x_{n-2,2} & x_{n-2,3} \\ y_{n-1} & x_{n-1,1} & \chi_{n-1,2}^{\ell-1} & x_{n-1,3} \\ \Psi_n^{\ell-1} & \chi_{n1}^{\ell-1} & \chi_{n2}^{\ell-1} & x_{n3} \end{array}\right] \quad \left[\begin{array}{cccc} & \text{Iteration } \ell & & \\ y_1 & \cdot & x_{12} & x_{13} \\ \cdot & r_{21} & x_{22} & \cdot \\ y_3 & x_{31} & x_{32} & x_{33} \\ \ldots & & & \\ r_{n-2} & x_{n-2,1} & x_{n-2,2} & x_{n-2,3} \\ y_{n-1} & x_{n-1,1} & \chi_{n-1,2}^{\ell} & x_{n-1,3} \\ \cdot & r_{n1} & \chi_{n2}^{\ell} & x_{n3} \end{array}\right]$$

Here, y indicates complete (non-missing) reports for the variable currently the subject of imputation; Ψ^p represents round p imputations of missing values of y; x represents complete reports of the of the set of variables available to condition the imputation; χ^p represents completed imputations of x from iteration p. Variables that were originally reported as a range but are not currently imputed are represented by r, and $\cdot$ represents values that are completely missing that are not yet imputed. Every x variable becomes a y variable at its place in the sequence of imputations within each iteration. Note that no missing values remain in the stylized $\ell - 1$ data set.

Ideally, one would like to condition every imputation on as many variables as possible, as well as on interactions and higher powers of those terms. Of course there are always practical limits to such a strategy due to degrees of freedom constraints, and some judgement must be applied in selecting a "maximal" set of conditioning variables, X. Of that maximal set, not every element may be non-missing at a given stage of imputation. For each variable to be imputed, the FRITZ algorithm determines the set of non-missing variables among the maximal set of of conditioning variables for each observation, denoted $X_{(i)}$ for observation i. Given the set of available conditioning variables $X_{(i)}$, the model essentially regresses the target imputation variable on the subset of conditioning variables using values *from the previous iteration of the model*. This process is made more efficient by estimating a maximal normalized cross-product matrix for each variable to be imputed, denoted $\sum (X, Y)_{\ell-1}$, and then subsetting the rows and columns corresponding to the non-missing conditioning variables for a given observation, denoted $\sum \left(X_{(i)}, Y\right)_{\ell-1}$. The imputation for observation i in iteration ℓ is thus given by:

$$\Psi_{i\ell} = \beta_{(i)\ell} X_{(i)i\ell} + e_{i\ell} \tag{25}$$

where $X_{(i)i\ell}$ is the rows of $X_{(i)\ell}$ corresponding to i; $X_{(i)\ell}$ is the subset of X that is available for i in iteration ℓ; $\beta_{(i)\ell} = \sum \left(X_{(i)} X_{(i)}\right)_{\ell-1}^{-1} \sum \left(X_{(i)} Y\right)_{\ell-1}$, and $e_{i\ell}$ is a random error term. Once a value is imputed, its imputed value is used (along with reported values) in conditioning later imputations.

The choice of error term $e_{i\ell}$ has been the subject of several experiments (see (Kennickell 1998)). In early releases of the SCF, $e_{i\ell}$ was taken to be a draw from a truncated normal distribution. The draw was restricted to the central 95 percent of the distribution, with occasional supplementary constraints imposed by the structure of the data or respondent-provided ranges for the variable under imputation. More recently, $e_{i\ell}$ has been drawn from an empirical distribution.

The FRITZ algorithm for imputing multinomial and binary variables works similarly, with an appropriate "regression" substituted for (25).

Using The FRITZ Algorithm for Disclosure Limitation The FRITZ algorithm is applied to the confidentiality protection problem in a straightforward manner. In the 1995 SCF, all dollar values for selected cases were simulated. The procedure is as follows. First, a set of cases which present excessive disclosure risk are selected (see (Kennickell 1997)). These are selected on the basis of having unusual levels of wealth or income given other characteristics, or other unusual combinations of responses. Second, a random set of cases is selected to reduce the ability of an intruder to determine even the set of cases

determined to present an excessive disclosure risk. Then, a new data set is created for all the selected cases, and shadow variables (which detail the "type" of response given for a particular case-variable pair, e.g., a complete report, a range report, or non-response) are set so that the FRITZ model interprets the responses as range responses. The type of range mimics one where the respondent volunteered a dollar range – a dollar amount of $\pm p$ percent (where p is and undisclosed number between 10 and 20 percent) is stored in a data set normally used to contain unique range reports. Finally, the actual dollar values are set to missing, and the FRITZ algorithm is applied to the selected cases, using the simulated range reports to constrain the imputed values. Subsequent evaluation of the 1995 SCF ((Fries, Johnson, and Woodburn 1997)) indicates that while the imputations substantially masked individual cases, the effect on important distributional characteristics was minimal.

A.3.3 Other Methods

(Moore 1996b) This paper presents a brief overview of data-swapping techniques for disclosure limitation, and presents a more sophisticated technique than found elsewhere in the literature. The author presents an algorithm for a controlled data swap based on the rank-based proximity swap of (Greenberg 1987). The contribution in this paper is to provide a technique which preserves univariate and bivariate relationships in the data. Based on a simulation using the 1993 Annual Housing Survey Public Use File, the author concludes that the algorithm preserves the desired moments to an acceptable degree (and hence retains some degree of analytic usefulness), while providing a level of confidentiality protection comparable to simple additive-noise methods.

(Moore 1996c) This paper suggests modifications to the Confidentiality Edit, the data-swapping procedure used for disclosure limitation in the 1990 Decennial Census. The suggested improvements are based on the ARGUS system for determining high-risk cases (see (Hundepool and Willenborg 1999) and (Nordholt 1999) below), and the German SAFE system for perturbing data. The author also presents two measures of the degree of distortion induced by the swap, and an algorithm to minimize this distortion.

(Mayda, Mohl, and Tambay 1997) This paper examines the relationship between variance estimation and confidentiality protection in surveys with complex designs. In particular, the authors consider the case of the Canadian National Population Health Survey (NPHS), a longitudinal survey with a multi-stage clustered design. To prepare a public use file, it was deemed necessary to remove specific design information such as stratum and cluster identifiers due to the extremely detailed level of geography they represented. Furthermore, providing cluster information could allow users to reconstitute households, increasing the probability of identifying individuals. However, specific design information is necessary to correctly compute variances using jackknife or other methods. This highlights yet another aspect of the conflict between providing high quality data and protecting confidentiality. The authors describe the approach taken to resolve this conflict. Specifically, strata and clusters are collapsed to form "super-strata" and "super-clusters" in the public use file, which protect confidentiality while providing enough information for researchers to obtain unbiased variance estimates under certain conditions. The drawback of this approach is that it does not generate the exact variance corresponding to the original design, and that collapsing reduces degrees of freedom and hence the precision of variance estimates.

(Nadeau, Gagnon, and Latouche 1999) This paper presents a discussion of confidentiality issues surrounding Statistics Canada's Survey of Labour and Income Dynamics (SLID), and presents the release strategy for microdata on individual and family income. SLID is a longitudinal survey designed to support studies of economic well-being of individuals and families, and of their determinants over time. With the demise of the Canadian Survey of Consumer Finances (SCF) in 1998, SLID became the official source

of information for both longitudinal *and* cross-sectional income data on individuals and families. This presented some rather unique issues for disclosure limitation.

Prior to integrating SLID and SCF, Statistics Canada did not release sufficient information in the SLID Public Use Microdata Files (PUMFs) to allow household reconstitution. It was considered too difficult to protect confidentiality at the household level in a longitudinal microdata file. However, since integrating SLID and SCF, it has become a priority to release cross-sectional PUMFs that meet needs of former SCF users. In particular, the cross-sectional PUMFs now contain household and family identifiers, which allow household and family reconstitution. This compromises the release of longitudinal PUMFs. Instead, Statistics Canada has opted to explore other options for the release of longitudinal data – namely release of synthetic files, and creation of research data centres. In the meantime, a number of disclosure limitation methods have been explored for the cross-sectional PUMFs to limit the ability of intruders to link records dynamically (constructing their own longitudinal file, considered "too risky" for re-identification) and/or re-identify records by linking to the Income Tax Data File (ITDF).

The disclosure control methods applied in the cross-sectional PUMFs include both data reduction and data modification methods. The data reduction methods include dropping direct identifiers, aggregating geographic variables, and categorical grouping for some occupational variables. Data modification methods are applied to numeric variables. In particular, year of birth is perturbed with additive noise; income variables are both bottom- and top-coded, and the remaining values are perturbed with a combined random-rounding and additive noise method.

Finally, the authors assess how successful these measures are at protecting confidentiality and maintaining analytical usefulness. To address the former, they consider both linking consecutive cross-sectional PUMFs and linking to the ITDF. In both cases, they consider both direct matches and nearest-neighbor matches. They find that the ability of an intruder to match records in either consecutive PUMFs or to the ITDF is severely limited by the disclosure control measures. As for the usefulness of the data, they find little difference in the marginal distribution of most variables at highly aggregated levels (*i.e.*, the national level), but more significant differences at lower levels of aggregation (i.e., the province×sex level).

(Hundepool and Willenborg 1999), (Nordholt 1999) These papers describe the τ-ARGUS and μ-ARGUS software packages developed by Statistics Netherlands for disclosure limitation. (Nordholt 1999) describes their specific application to the Annual Survey on Employment and Earnings (ASEE). The τ-ARGUS software tackles the problem of disclosure limitation in tabular data. It automatically applies a series of primary and secondary suppressions to tabular data on the basis of a dominance rule: a cell is considered unsafe if the n major contributors to that cell are responsible for at least p percent of the total cell value. The μ-ARGUS software is used to create a public use microdata file from the ASEE. The public use microdata has to satisfy two criteria, which are implemented with μ-ARGUS: first, every category of an identifying variable must occur "frequently enough" (200,000 times is the default for ASEE); second, every bivariate combination of values must occur "frequently enough" (1,000 times is the default for ASEE). These objectives are achieved via global recoding and local suppression.

A.4 Analysis of Disclosure-Proofed Data

(Little 1993) (Little 1993) develops a model-based likelihood theory for the analysis of masked data. His approach is to formally model the mechanism whereby case-variable pairs are selected for masking, the masking method, and derive an appropriate model for analysis of the resulting data. His method is sufficiently general to allow for a variety of masking selection mechanisms, and such diverse masking methods as deletion, coarsening, imputation, and aggregation. The formal theory follows.

Let $\mathbf{X} = \{x_{ij}\}$ denote an $(n \times p)$ unmasked data matrix of n observations on p variables. Let $\mathbf{M} = \{m_{ij}\}$ denote the masking indicator matrix, where $m_{ij} = 1$ if x_{ij} is masked, and $m_{ij} = 0$ otherwise. Let $\mathbf{Z} = \{z_{ij}\}$

denote the masked data, i.e., z_{ij} is the masked value of x_{ij} if $m_{ij} = 1$, and $z_{ij} = x_{ij}$ if $m_{ij} = 0$. Model the joint distribution of $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{M}$ with the density function:

$$f(\mathbf{X}, \mathbf{Z}, \mathbf{M}|\boldsymbol{\theta}) = f_X(\mathbf{X}|\boldsymbol{\theta}) f_Z(\mathbf{Z}|\mathbf{X}) f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}). \tag{26}$$

Here $f_X(\mathbf{X}|\boldsymbol{\theta})$ is the density of the unmasked data given unknown parameters $\boldsymbol{\theta}$, which would be the basis for analysis in the absence of masking; $f_Z(\mathbf{Z}|\mathbf{X})$ formalizes the masking treatment; and $f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z})$ formalizes the masking selection mechanism. If the analyst knows which values are masked and the method masking, then the analyst knows $\mathbf{M}$, as well as the distributions of $\mathbf{M}$ and $\mathbf{Z}$. If not, then $\mathbf{M}$ is unknown. A more general specification would also index the distributions of $\mathbf{M}$ and/or $\mathbf{Z}$ by unknown parameters, and a full likelihood analysis would then involve both $\boldsymbol{\theta}$ and these unknown masking parameters.

Let $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$ and $\mathbf{Z} = (\mathbf{Z}_{obs}, \mathbf{Z}_{mis})$ where *obs* denotes observed components, and *mis* denotes missing components of each matrix. Analysis of the masked data is based on the likelihood for $\boldsymbol{\theta}$ given the data $\mathbf{M}, \mathbf{X}_{obs}$, and $\mathbf{Z}_{obs}$. This is obtained formally by integrating the joint density in (26) over the missing values $\mathbf{X}_{mis}$ and $\mathbf{Z}_{mis}$:

$$L(\boldsymbol{\theta}|\mathbf{M}, \mathbf{X}_{obs}, \mathbf{Z}_{obs}) = \int f_X(\mathbf{X}|\boldsymbol{\theta}) f_Z(\mathbf{Z}|\mathbf{X}) f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}) d\mathbf{X}_{mis} d\mathbf{Z}_{mis}. \tag{27}$$

Since the distribution of $\mathbf{M}$ in (27) may depend on $\mathbf{X}$ and $\mathbf{Z}_{obs}$, but should not depend on $\mathbf{Z}_{mis}$, we can write $f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}) = f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}_{obs})$. Thus we can rewrite (27) as:

$$L(\boldsymbol{\theta}|\mathbf{M}, \mathbf{X}_{obs}, \mathbf{Z}_{obs}) = \int f_X(\mathbf{X}|\boldsymbol{\theta}) f_Z^*(\mathbf{Z}_{obs}|\mathbf{X}) f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}_{obs}) d\mathbf{X}_{mis} \tag{28}$$

where $f_Z^*(\mathbf{Z}_{obs}|\mathbf{X}) = \int f_Z(\mathbf{Z}|\mathbf{X}) d\mathbf{Z}_{mis}$.

The author notes that the likelihood in (28) can be simplified if the masking selection and treatment mechanisms satisfy certain ignorability conditions, in the sense of (Rubin 1976) and (Rubin 1978). Specifically, if the masking selection mechanism is ignorable, then $f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}) = f_M(\mathbf{M}|\mathbf{X}_{obs}, \mathbf{Z}_{obs})$ for all $\mathbf{X}_{mis}$, $\mathbf{Z}_{mis}$. In this case, the density of $\mathbf{M}$ can be omitted from (28). Similarly, the masking treatment mechanism is ignorable if $f_Z^*(\mathbf{Z}_{obs}|\mathbf{X}) = f_Z^*(\mathbf{Z}_{obs}|\mathbf{X}_{obs})$ for all $\mathbf{X}_{mis}$. In this case, the density of $\mathbf{Z}_{obs}$ can be omitted from (28). Finally, if both mechanisms are ignorable, then the likelihood reduces to:

$$L(\boldsymbol{\theta}|\mathbf{M}, \mathbf{X}_{obs}, \mathbf{Z}_{obs}) = \int f_X(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{X}_{mis}$$

which is proportional to the marginal density of $\mathbf{X}_{obs}$.

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999, March). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Bethlehem, J., W. Keller, and J. Pannekoek (1990). Disclosure control of microdata. *Journal of the American Statistical Association 85*, 38–45.

Boudreau, J.-R. (1995, November). Assessment and reduction of disclosure risk in microdata files containing discrete data. Presented at Statistics Canada Symposium 95.

Duncan, G. T. and S. Mukherjee (1998, June). Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise. Heinz School of Public Policy and Management Working Paper No. 1998-15.

Evans, B. T., R. Moore, and L. Zayatz (1996). New directions in disclosure limitation at the Census Bureau. U.S. Census Bureau Research Report No. LVZ96/01.

Evans, T., L. Zayatz, and J. Slanta (1998). Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistics 14*(4), 537–551.

Fienberg, S. E. (1994, December). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Carnegie Mellon University Department of Statistics Technical Report No. 611.

Fienberg, S. E. (1997, September). Confidentiality and disclosure limitation methodology: Challenges for national statistics and statistical research. Presented at the Committee on National Statistics 25th Anniversary Meeting. Carnegie Mellon Department of Statistics Technical Report, Working Paper No. 668.

Fienberg, S. E. and U. E. Makov (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics 14*(4), 385–397.

Fienberg, S. E., U. E. Makov, and R. J. Steele (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics 14*(4), 485–502.

Franconi, L. (1999, March). Level of safety in microdata: Comparisons between different definitions of disclosure risk and estimation models. Presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, Greece, Working Paper No. 4.

Fries, G., B. W. Johnson, and R. L. Woodburn (1997, September). Analyzing disclosure review procedures for the Survey of Consumer Finances. SCF Working Paper, presented at the 1997 Joint Statistical Meetings, Anaheim, CA.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics 9*(2), 383–406.

Greenberg, B. (1987). Rank swapping for masking ordinal microdata. U.S. Census Bureau, unpublished manuscript.

Hundepool, A. and L. Willenborg (1999, March). ARGUS: Software from the SDC project. Presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, Greece, Working Paper No. 7.

Jabine, T. B. (1993). Statistical disclosure limitation practices of the United States statistical agencies. *Journal of Official Statistics 9*(2), 427–454.

Kennickell, A. B. (1991, October). Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. SCF Working Paper, prepared for the Annual Meetings of the Amercian Statistical Association, Atlanta, Georgia, August 1991.

Kennickell, A. B. (1997, November). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. SCF Working Paper.

Kennickell, A. B. (1998, September). Multiple imputation in the Survey of Consumer Finances. SCF Working Paper, prepared for the August 1998 Joint Statistical Meetings, Dallas, TX.

Kennickell, A. B. (2000, May). Wealth measurement in the Survey of Consumer Finances: Methodology and directions for future research. SCF Working Paper, prepared for the May 2000 annual meetings of the American Association for Public Opinion Research, Portland, Oregon.

Kim, J. J. and W. E. Winkler (1997). Masking microdata files. U.S. Census Bureau Research Report No. RR97/03.

Kooiman, P. (1998). Comment on disclosure limitation for categorical data. *Journal of Official Statistics 14*(4), 503–508.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics 9*(2), 313–331.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics 9*(2), 407–426.

Little, R. J. and D. B. Rubin (1987). *Statistical Analysis of Missing Data*. New York: Wiley.

Mayda, J., C. Mohl, and J. Tambay (1997). Variance estimation and confidentiality: They are related! Unpublished Manuscript, Statistics Canada.

Moore, Jr, R. A. (1996a). Analysis of the Kim-Winkler algorithm for masking microdata files - how much masking is necessary and sufficient? Conjectures for the development of a controllable algorithm. U.S. Census Bureau Research Report No. RR96/05.

Moore, Jr, R. A. (1996b). Controlled data-swapping techniques for masking public use microdata sets. U.S. Census Bureau Research Report No. RR96/04.

Moore, Jr, R. A. (1996c). Preliminary recommendations for disclosure limitation for the 2000 Census: Improving the 1990 confidentiality edit procedure. U.S. Census Bureau Statstical Research Report Series, No. RR96/06.

Nadeau, C., E. Gagnon, and M. Latouche (1999). Disclosure control strategy for the release of microdata in the Canadian Survey of Labour and Income Dynamics. Presented at the 1999 Joint Statistical Meetings, Baltimore, MD.

National Research Council (2000). Improving access to and confidentiality of research data: Report of a workshop. Committee on National Statistics, Christopher Mackie and Norman Bradburn, Eds. Commission on Behavioral and Social Sciences and Education, National Academy Press, Washington, D.C.

Nordholt, E. S. (1999, March). Statistical disclosure control of the Statistics Netherlands employment and earnings data. Presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, Greece, Working Paper No. 2.

Pursey, S. (1999, March). Disclosure control methods in the public release of a microdata file of small businesses. Presented at the Joint ECE/Eurostat Work Session on Statisical Data Confidentiality, Thessaloniki, Greece, Working Paper No. 5.

Raghunathan, T. E., J. M. Lepkowski, J. V. Hoewyk, and P. Solenberger (1998). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Research Center, University of Michigan.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics 6*, 34–58.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. (1993). Discussion of statistical disclosure limitation. *Journal of Official Statistics 9*(2), 461–468.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Skinner, C. and D. Holmes (1993). Modelling population uniqueness. In *Proceedings of International Seminar on Statistical Confidentiality*, Luxembourg, pp. 175–199. EUROSTAT.

Subcommittee on Disclosure Avoidance Techniques (1978a). Statistical policy working paper no. 2: Report on statistical disclosure and disclosure avoidance techniques. Federal Committee on Statistical Methodology, Office of Federal Policy and Standards, U.S. Department of Commerce, Washington, D.C.

Subcommittee on Disclosure Avoidance Techniques (1994b). Statistical policy working paper no. 22: Report on statistical disclosure limitation methodology. Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, U.S. Office of Management and Budget, Washington, D.C.

Willenborg, L. and J. Kardaun (1999, March). Fingerprints in microdata sets. Presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, Greece, Working Paper No. 10.

Winkler, W. E. (1997). Views on the production and use of confidential microdata. U.S. Census Bureau Research Report No. RR97/01.

Winkler, W. E. (1998). Producing public-use files that are analytically valid and confidential. U.S. Census Bureau Research Report No. RR98/02.

Figure 1

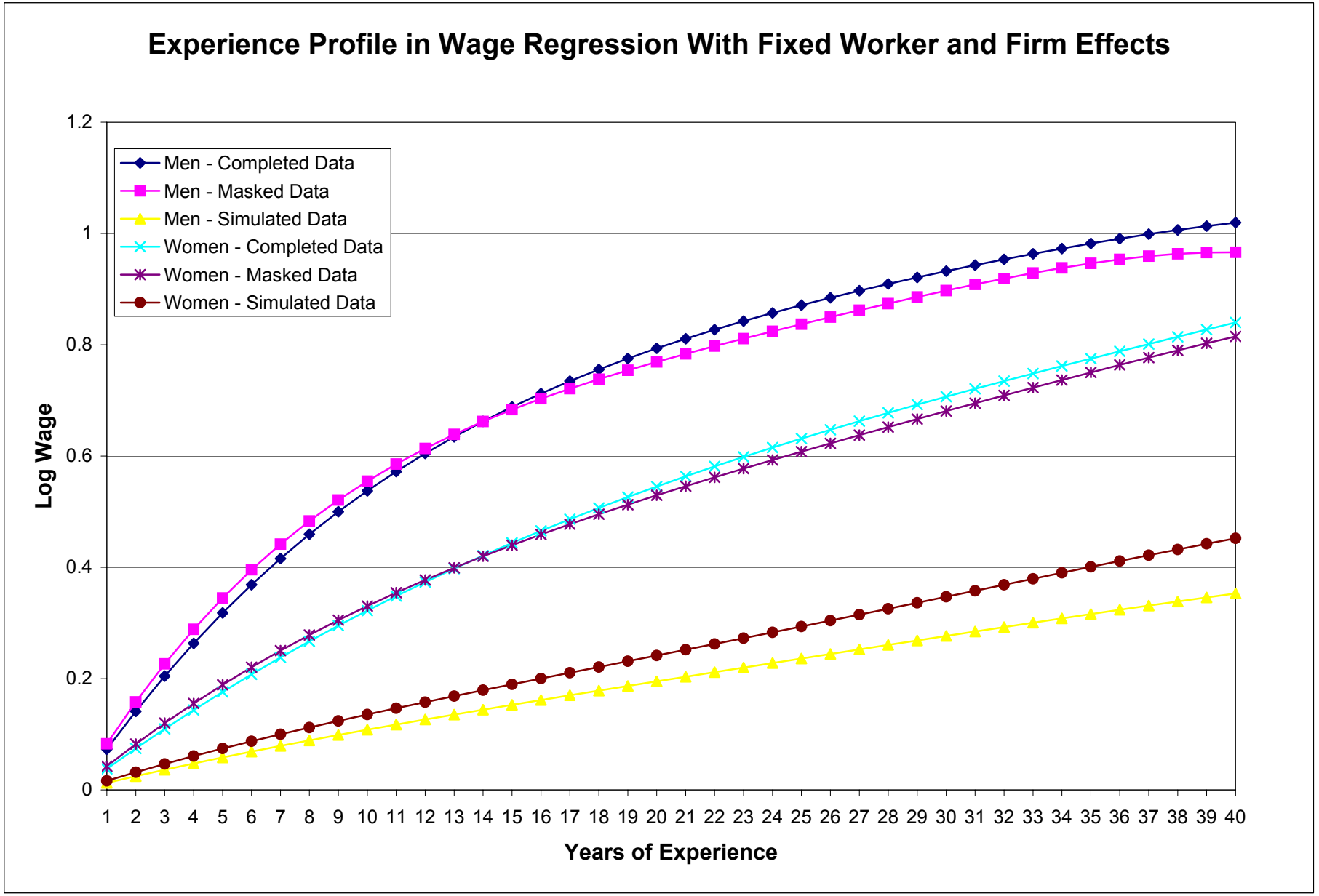
Experience Profile in Wage Regression With Fixed Worker and Firm Effects
Men - Completed Data
Men - Masked Data
Men - Simulated Data
Women - Completed Data
Women - Masked Data
Women - Simulated Data
Log Wage
0
0.2
0.4
0.6
0.8
1
1.2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Years of Experience

TABLE 1: UNIVARIATE STATISTICS ON COMPLETED, MASKED, AND SIMULATED DATA: MEN

Variable	N	Average Mean or Proportion in Category	Average Variance of Mean	Between Implicate Variance of Mean	Total Variance of Mean	Relative Bias	Relative Increase in Variance	Average Variance	Average Variance of Variance	Between Implicate Variance of Variance	Total Variance of Variance	Relative Bias	Relative Increase in Variance
COMPLETED DATA													
Year of Birth	201,906	1952	0.001	0	0.001			250.2	0.41	0	0.41		
No Diploma	201,906	0.299	1.03E-06	2.68E-03	2.95E-03			0.207	1.69E-07	3.78E-04	4.16E-04		
Elementary School	201,906	0.180	7.31E-07	7.83E-05	8.68E-05			0.148	2.99E-07	3.26E-05	3.62E-05		
Middle School	201,906	0.112	4.90E-07	4.09E-04	4.51E-04			0.099	2.92E-07	2.42E-04	2.67E-04		
High School	201,906	0.060	2.77E-07	4.67E-05	5.17E-05			0.056	2.14E-07	3.69E-05	4.08E-05		
Basic Vocational School	201,906	0.215	8.32E-07	6.00E-04	6.61E-04			0.168	2.69E-07	1.98E-04	2.18E-04		
Advanced Vocational School	201,906	0.060	2.81E-07	6.13E-05	6.78E-05			0.057	2.16E-07	4.62E-05	5.10E-05		
Technical College or University	201,906	0.036	1.71E-07	4.57E-05	5.04E-05			0.034	1.46E-07	3.85E-05	4.25E-05		
Graduate School	201,906	0.039	1.84E-07	8.16E-06	9.16E-06			0.037	1.57E-07	6.96E-06	7.81E-06		
Log Real Annual Compensation (1980 FF 000)	1,893,555	4.164	4.70E-07	2.37E-09	4.73E-07			0.891	4.19E-06	8.79E-09	4.20E-06		
Days Paid (max 360)	1,893,555	263.0	0.009	0	0.009			17987	180	0	180		
MASKED DATA													
Year of Birth	201,906	1951	0.001	0.009	0.011	-0.001	7.856	236.4	0.43	1.75	2.36	-0.055	4.734
No Diploma	201,906	0.309	1.03E-06	5.32E-03	5.86E-03	0.033	0.982	0.209	1.61E-07	5.35E-04	5.89E-04	0.007	0.414
Elementary School	201,906	0.181	7.31E-07	8.90E-04	9.80E-04	0.005	10.284	0.148	2.93E-07	3.72E-04	4.09E-04	-0.001	10.319
Middle School	201,906	0.105	4.63E-07	6.16E-04	6.78E-04	-0.061	0.503	0.093	2.83E-07	3.90E-04	4.29E-04	-0.056	0.608
High School	201,906	0.056	2.63E-07	7.66E-05	8.46E-05	-0.054	0.636	0.053	2.06E-07	5.98E-05	6.60E-05	-0.051	0.620
Basic Vocational School	201,906	0.208	8.12E-07	9.62E-04	1.06E-03	-0.031	0.602	0.164	2.73E-07	3.44E-04	3.79E-04	-0.024	0.737
Advanced Vocational School	201,906	0.070	3.22E-07	2.82E-04	3.10E-04	0.163	3.579	0.065	2.35E-07	2.07E-04	2.28E-04	0.147	3.468
Technical College or University	201,906	0.032	1.53E-07	6.41E-05	7.07E-05	-0.105	0.402	0.031	1.33E-07	5.54E-05	6.11E-05	-0.103	0.438
Graduate School	201,906	0.038	1.83E-07	1.13E-05	1.26E-05	-0.009	0.376	0.037	1.56E-07	9.66E-06	1.08E-05	-0.008	0.380
Log Real Annual Compensation (1980 FF 000)	1,893,555	4.146	4.58E-07	4.97E-06	5.92E-06	-0.004	11.520	0.867	3.87E-06	4.07E-06	8.34E-06	-0.027	0.987
Days Paid (max 360)	1,893,555	257.4	0.009	0.058	0.073	-0.021	6.672	17039	166	1080	1353	-0.053	6.519
SIMULATED DATA													
Year of Birth	201,906	1951	0.001	0.157	0.173	0.000	138.920	174.8	0.24	1.41	1.79	-0.301	3.359
No Diploma	201,906	0.347	1.08E-06	1.02E-02	1.12E-02	0.160	2.798	0.217	1.32E-07	4.60E-04	5.06E-04	0.049	0.215
Elementary School	201,906	0.142	6.02E-07	6.88E-04	7.57E-04	-0.209	7.722	0.122	3.03E-07	3.47E-04	3.82E-04	-0.177	9.573
Middle School	201,906	0.107	4.72E-07	5.90E-04	6.49E-04	-0.041	0.440	0.095	2.86E-07	3.59E-04	3.96E-04	-0.038	0.483
High School	201,906	0.059	2.72E-07	2.77E-04	3.05E-04	-0.016	4.910	0.055	2.08E-07	2.17E-04	2.39E-04	-0.019	4.866
Basic Vocational School	201,906	0.217	8.26E-07	3.39E-03	3.73E-03	0.012	4.651	0.167	2.56E-07	1.04E-03	1.14E-03	-0.007	4.247
Advanced Vocational School	201,906	0.064	2.95E-07	2.94E-04	3.24E-04	0.056	3.775	0.060	2.21E-07	2.16E-04	2.38E-04	0.049	3.653
Technical College or University	201,906	0.031	1.49E-07	8.54E-05	9.41E-05	-0.133	0.866	0.030	1.29E-07	7.42E-05	8.18E-05	-0.129	0.924
Graduate School	201,906	0.033	1.58E-07	2.87E-05	3.17E-05	-0.144	2.458	0.032	1.38E-07	2.52E-05	2.78E-05	-0.140	2.559
Log Real Annual Compensation (1980 FF 000)	1,893,555	4.172	4.17E-07	2.20E-04	2.43E-04	0.002	512.000	0.789	2.00E-06	5.93E-04	6.54E-04	-0.115	154.885
Days Paid (max 360)	1,893,555	256.5	0.007	2.690	2.967	-0.025	311.300	13990	144	41251	45521	-0.222	251.871

Notes: Education categories are the highest degree attained. Relative bias and variance are computed in comparison to the completed data.
Sources: Authors' calculations based upon the INSEE DADS and EDP data 1976-1996.

TABLE 2: UNIVARIATE STATISTICS ON COMPLETED, MASKED, AND SIMULATED DATA: WOMEN

VARIABLE	N	Average Mean or Proportion in Category	Average Variance of Mean	Between Implicate Variance of Mean	Total Variance of Mean	Relative Bias	Relative Increase in Variance	Average Variance	Average Variance of Variance	Between Implicate Variance of Variance	Total Variance of Variance	Relative Bias	Relative Increase in Variance
COMPLETED DATA													
Year of Birth	161,007	1954	0.001	0	0.001			224.7	0.52	0	0.52		
No Diploma	161,007	0.264	1.19E-06	2.74E-03	3.01E-03			0.192	2.66E-07	5.47E-04	6.02E-04		
Elementary School	161,007	0.200	9.93E-07	1.17E-04	1.30E-04			0.160	3.57E-07	4.24E-05	4.70E-05		
Middle School	161,007	0.150	7.90E-07	3.84E-04	4.24E-04			0.127	3.84E-07	1.89E-04	2.08E-04		
High School	161,007	0.082	4.65E-07	8.58E-05	9.49E-05			0.075	3.24E-07	5.92E-05	6.54E-05		
Basic Vocational School	161,007	0.148	7.81E-07	1.90E-04	2.10E-04			0.126	3.86E-07	9.62E-05	1.06E-04		
Advanced Vocational School	161,007	0.074	4.27E-07	8.23E-05	9.10E-05			0.069	3.08E-07	5.85E-05	6.47E-05		
Technical College or University	161,007	0.058	3.37E-07	7.89E-05	8.71E-05			0.054	2.63E-07	6.03E-05	6.66E-05		
Graduate School	161,007	0.025	1.50E-07	3.14E-06	3.60E-06			0.024	1.35E-07	2.83E-06	3.25E-06		
Log Real Annual Compensation (1980 FF 000)	1,319,819	3.777	8.11E-07	4.32E-09	8.16E-07			1.071	6.39E-06	1.65E-08	6.40E-06		
Days Paid (max 360)	1,319,819	260.9	0.014	0	0.014			18122	245	0	245		
MASKED DATA													
Year of Birth	161,007	1953	0.001	0.016	0.018	0.000	12.245	212.1	0.50	3.36	4.20	-0.056	7.089
No Diploma	161,007	0.239	1.09E-06	6.79E-03	7.47E-03	-0.094	1.482	0.176	3.01E-07	1.05E-03	1.15E-03	-0.084	0.913
Elementary School	161,007	0.193	9.65E-07	6.09E-04	6.70E-04	-0.034	4.147	0.155	3.60E-07	2.33E-04	2.57E-04	-0.029	4.462
Middle School	161,007	0.161	8.31E-07	1.55E-03	1.71E-03	0.074	3.034	0.134	3.71E-07	6.82E-04	7.50E-04	0.052	2.601
High School	161,007	0.082	4.67E-07	1.25E-04	1.38E-04	0.006	0.457	0.075	3.25E-07	8.94E-05	9.87E-05	0.005	0.508
Basic Vocational School	161,007	0.164	8.43E-07	1.26E-03	1.39E-03	0.107	5.616	0.136	3.72E-07	5.89E-04	6.49E-04	0.079	5.109
Advanced Vocational School	161,007	0.073	4.21E-07	2.85E-04	3.14E-04	-0.013	2.449	0.068	3.02E-07	2.03E-04	2.24E-04	-0.015	2.458
Technical College or University	161,007	0.063	3.66E-07	1.69E-04	1.86E-04	0.094	1.139	0.059	2.77E-07	1.29E-04	1.42E-04	0.086	1.136
Graduate School	161,007	0.025	1.50E-07	3.16E-06	3.63E-06	0.001	0.008	0.024	1.36E-07	2.87E-06	3.29E-06	0.001	0.013
Log Real Annual Compensation (1980 FF 000)	1,319,819	3.760	7.89E-07	1.38E-06	2.31E-06	-0.005	1.828	1.042	5.83E-06	3.95E-06	1.02E-05	-0.027	0.588
Days Paid (max 360)	1,319,819	254.5	0.013	0.089	0.111	-0.024	7.055	17010	222	1188	1529	-0.061	5.254
SIMULATED DATA													
Year of Birth	161,007	1954	0.001	0.231	0.255	0.000	181.465	160.0	0.28	1.62	2.06	-0.288	2.970
No Diploma	161,007	0.262	1.15E-06	8.66E-03	9.53E-03	-0.007	2.163	0.185	2.69E-07	1.27E-03	1.40E-03	-0.032	1.324
Elementary School	161,007	0.172	8.80E-07	4.92E-04	5.42E-04	-0.143	3.162	0.142	3.76E-07	2.25E-04	2.48E-04	-0.114	4.265
Middle School	161,007	0.162	8.36E-07	1.05E-03	1.16E-03	0.077	1.731	0.135	3.75E-07	4.84E-04	5.33E-04	0.058	1.557
High School	161,007	0.089	4.99E-07	3.39E-04	3.73E-04	0.085	2.936	0.080	3.34E-07	2.28E-04	2.51E-04	0.074	2.840
Basic Vocational School	161,007	0.158	8.22E-07	7.62E-04	8.39E-04	0.069	2.988	0.132	3.78E-07	3.75E-04	4.13E-04	0.052	2.891
Advanced Vocational School	161,007	0.082	4.63E-07	3.42E-04	3.76E-04	0.097	3.134	0.075	3.20E-07	2.38E-04	2.63E-04	0.085	3.058
Technical College or University	161,007	0.053	3.08E-07	1.74E-04	1.92E-04	-0.090	1.201	0.050	2.44E-07	1.37E-04	1.51E-04	-0.087	1.267
Graduate School	161,007	0.024	1.48E-07	1.24E-05	1.38E-05	-0.014	2.833	0.024	1.34E-07	1.13E-05	1.26E-05	-0.014	2.866
Log Real Annual Compensation (1980 FF 000)	1,319,819	3.789	7.41E-07	1.59E-03	1.75E-03	0.003	2140.856	0.978	3.01E-06	3.92E-03	4.32E-03	-0.087	673.020
Days Paid (max 360)	1,319,819	253.3	0.010	5.795	6.384	-0.029	463.974	13327	185	41295	45609	-0.265	185.503

Notes: Education categories are the highest degree attained. Relative bias and variance are computed in comparison to the completed data.
Sources: Authors' calculations based upon the INSEE DADS and EDP data 1976-1996.

TABLE 3: UNIVARIATE STATISTICS ON COMPLETED, MASKED, AND SIMULATED DATA: FIRMS

VARIABLE	N	Average Mean	Average Variance of Mean	Between Implicate Variance of Mean	Total Variance of Mean	Relative Bias	Relative Increase in Variance	Average Variance	Average Variance of Variance	Between Implicate Variance of Variance	Total Variance of Variance	Relative Bias	Relative Increase in Variance
COMPLETED DATA													
Log Sales (FF millions)	470,812	10.58	4.39E-06	3.19E-05	3.95E-05			2.07	2.59E-05	4.95E-05	8.04E-05		
Log Capital Stock (FF millions)	470,812	8.50	9.51E-06	6.86E-06	1.71E-05			4.48	1.32E-04	6.28E-05	2.01E-04		
Log Average Employment	470,812	4.33	2.57E-06	9.44E-05	1.06E-04			1.21	1.20E-05	7.45E-05	9.40E-05		
MASKED DATA													
Log Sales (FF millions)	470,812	10.56	4.37E-06	2.54E-05	3.23E-05	-0.002	-0.183	2.06	2.56E-05	4.32E-05	7.30E-05	-0.012	-0.091
Log Capital Stock (FF millions)	470,812	8.48	9.45E-06	1.47E-05	2.56E-05	-0.002	0.502	4.45	1.28E-04	6.23E-05	1.96E-04	-0.028	-0.021
Log Average Employment	470,812	4.31	2.55E-06	8.93E-05	1.01E-04	-0.004	-0.054	1.20	1.17E-05	7.85E-05	9.81E-05	-0.024	0.044
SIMULATED DATA													
Log Sales (FF millions)	470,378	10.59	1.45E-06	2.30E-04	2.54E-04	0.001	5.439	0.68	2.16E-06	1.22E-04	1.37E-04	0.704	0.704
Log Capital Stock (FF millions)	470,378	8.51	3.79E-06	2.52E-03	2.77E-03	0.002	161.704	1.78	1.93E-05	1.83E-03	2.04E-03	9.154	9.154
Log Average Employment	470,378	4.34	7.26E-07	5.55E-04	6.11E-04	0.002	4.739	0.34	8.70E-07	2.83E-04	3.12E-04	2.318	2.318

Notes: Relative bias and variance are computed in comparison to the completed data. Unweighted statistics.
Sources: Authors' calculations based on the INSEE EAE data 1978-1996.

TABLE 4: BIVARIATE STATISTICS FOR INDIVIDUALS: MEN AND WOMEN COMBINED

CORRELATIONS IN COMPLETED DATA (Between Implicate Variance Above Diagonal)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Male	1	*0.00000*	*0.00320*	*0.00016*	*0.00135*	*0.00071*	*0.00081*	*0.00015*	*0.00082*	*0.00008*
(2) Year of Birth	-0.07	1	*0.00029*	*0.00008*	*0.00038*	*0.00006*	*0.00003*	*0.00004*	*0.00003*	*0.00001*
(3) No Diploma	0.04	0.15	1	*0.00071*	*0.00017*	*0.00019*	*0.00036*	*0.00035*	*0.00018*	*0.00011*
(4) Elementary School	-0.03	-0.34	-0.30	1	*0.00030*	*0.00006*	*0.00033*	*0.00007*	*0.00004*	*0.00002*
(5) Middle School	-0.06	0.18	-0.24	-0.19	1	*0.00011*	*0.00043*	*0.00010*	*0.00007*	*0.00004*
(6) High School	-0.04	0.09	-0.17	-0.13	-0.10	1	*0.00012*	*0.00003*	*0.00002*	*0.00001*
(7) Basic Vocational School	0.08	-0.02	-0.30	-0.23	-0.18	-0.13	1	*0.00011*	*0.00008*	*0.00005*
(8) Advanced Vocational School	-0.03	-0.02	-0.17	-0.13	-0.10	-0.07	-0.13	1	*0.00002*	*0.00001*
(9) Technical College or University	-0.05	0.02	-0.14	-0.11	-0.08	-0.06	-0.10	-0.06	1	*0.00001*
(10) Graduate School	0.04	-0.06	-0.12	-0.09	-0.07	-0.05	-0.09	-0.05	-0.04	1

CORRELATIONS IN MASKED DATA (Between Implicate Variance Above Diagonal)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Male	1	*0.00001*	*0.01511*	*0.00340*	*0.00204*	*0.00052*	*0.00470*	*0.00268*	*0.00114*	*0.00011*
(2) Year of Birth	-0.07	1	*0.00110*	*0.00014*	*0.00074*	*0.00013*	*0.00008*	*0.00004*	*0.00007*	*0.00001*
(3) No Diploma	0.08	0.13	1	*0.00121*	*0.00030*	*0.00028*	*0.00117*	*0.00052*	*0.00026*	*0.00014*
(4) Elementary School	-0.02	-0.30	-0.30	1	*0.00064*	*0.00012*	*0.00029*	*0.00027*	*0.00014*	*0.00004*
(5) Middle School	-0.08	0.17	-0.24	-0.18	1	*0.00032*	*0.00083*	*0.00020*	*0.00022*	*0.00011*
(6) High School	-0.05	0.08	-0.17	-0.13	-0.10	1	*0.00017*	*0.00005*	*0.00004*	*0.00002*
(7) Basic Vocational School	0.06	-0.02	-0.30	-0.23	-0.19	-0.13	1	*0.00012*	*0.00012*	*0.00005*
(8) Advanced Vocational School	-0.01	-0.02	-0.17	-0.13	-0.11	-0.07	-0.13	1	*0.00006*	*0.00002*
(9) Technical College or University	-0.07	0.02	-0.14	-0.10	-0.08	-0.06	-0.11	-0.06	1	*0.00002*
(10) Graduate School	0.04	-0.05	-0.11	-0.09	-0.07	-0.05	-0.09	-0.05	-0.04	1

CORRELATIONS IN SIMULATED DATA (Between Implicate Variance Above Diagonal)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Male	1	*0.00038*	*0.02479*	*0.00319*	*0.00360*	*0.00286*	*0.00578*	*0.00311*	*0.00212*	*0.00041*
(2) Year of Birth	-0.08	1	*0.00067*	*0.00010*	*0.00006*	*0.00015*	*0.00019*	*0.00015*	*0.00014*	*0.00003*
(3) No Diploma	0.09	0.02	1	*0.00132*	*0.00061*	*0.00040*	*0.00031*	*0.00106*	*0.00053*	*0.00013*
(4) Elementary School	-0.04	0.01	-0.29	1	*0.00029*	*0.00013*	*0.00085*	*0.00013*	*0.00006*	*0.00005*
(5) Middle School	-0.08	0.01	-0.26	-0.17	1	*0.00025*	*0.00116*	*0.00010*	*0.00010*	*0.00007*
(6) High School	-0.06	-0.02	-0.18	-0.12	-0.11	1	*0.00053*	*0.00007*	*0.00003*	*0.00003*
(7) Basic Vocational School	0.07	0.01	-0.32	-0.21	-0.19	-0.14	1	*0.00021*	*0.00018*	*0.00020*
(8) Advanced Vocational School	-0.03	0.00	-0.19	-0.12	-0.11	-0.08	-0.13	1	*0.00003*	*0.00002*
(9) Technical College or University	-0.05	-0.02	-0.14	-0.09	-0.08	-0.06	-0.10	-0.06	1	*0.00001*
(10) Graduate School	0.03	-0.05	-0.12	-0.07	-0.07	-0.05	-0.08	-0.05	-0.04	1

Notes: N=362,913.

Sources: Authors' calculations based on the INSEE DADS and EDP data.

TABLE 5: BIVARIATE STATISTICS BASED ON THE WORK HISTORY FILE: MEN AND WOMEN COMBINED

CORRELATIONS IN COMPLETED DATA (Between Implicate Variance Above Diagonal)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
(1) Male	1									*0.00330*	*0.00018*	*0.00126*	*0.00079*	*0.00107*	*0.00019*	*0.00109*	*0.00010*		*0.00000*	*0.00000*	*0.00000*	*0.00000*
(2) Full Time Employee	0.167	1								*0.00008*	*0.00001*	*0.00004*	*0.00002*	*0.00004*	*0.00000*	*0.00003*	*0.00001*		*0.00000*	*0.00000*	*0.00000*	*0.00004*
(3) Engineer, Professional, or Manager	0.097	-0.008	1							*0.00011*	*0.00004*	*0.00004*	*0.00007*	*0.00006*	*0.00006*	*0.00013*	*0.00009*		*0.00000*	*0.00000*	*0.00000*	*0.00000*
(4) Technician or Technical White Collar	0.001	0.031	-0.155	1						*0.00006*	*0.00004*	*0.00003*	*0.00001*	*0.00004*	*0.00004*	*0.00001*	*0.00000*		*0.00000*	*0.00000*	*0.00000*	*0.00000*
(5) Other White Collar	-0.380	-0.088	-0.205	-0.309	1					*0.00037*	*0.00006*	*0.00017*	*0.00011*	*0.00024*	*0.00004*	*0.00011*	*0.00001*		*0.00000*	*0.00000*	*0.00000*	*0.00001*
(6) Skilled Blue Collar	0.314	0.132	-0.176	-0.265	-0.351	1				*0.00049*	*0.00007*	*0.00006*	*0.00002*	*0.00035*	*0.00003*	*0.00003*	*0.00000*		*0.00000*	*0.00000*	*0.00000*	*0.00001*
(7) Unskilled Blue Collar	0.030	-0.063	-0.158	-0.238	-0.316	-0.271	1			*0.00006*	*0.00007*	*0.00005*	*0.00000*	*0.00007*	*0.00001*	*0.00001*	*0.00000*		*0.00000*	*0.00000*	*0.00000*	*0.00000*
(8) Works in Ile-de-France	-0.018	-0.007	0.133	0.066	0.034	-0.092	-0.104	1		*0.00003*	*0.00000*	*0.00001*	*0.00001*	*0.00001*	*0.00001*	*0.00002*	*0.00003*		*0.00000*	*0.00000*	*0.00000*	*0.00000*
(9) Year of Birth	-0.070	-0.106	-0.094	-0.053	0.082	-0.053	0.083	-0.037	1	*0.00019*	*0.00009*	*0.00031*	*0.00005*	*0.00001*	*0.00002*	*0.00003*	*0.00001*		*0.00000*	*0.00000*	*0.00000*	*0.00006*
(10) No Diploma	0.043	-0.046	-0.102	-0.125	-0.039	0.072	0.166	-0.018	0.086	1	*0.00089*	*0.00023*	*0.00024*	*0.00060*	*0.00039*	*0.00024*	*0.00013*		*0.00009*	*0.00021*	*0.00016*	*0.00020*
(11) Elementary School	-0.021	0.028	-0.087	-0.063	0.006	0.073	0.041	-0.033	-0.286	-0.269	1	*0.00030*	*0.00006*	*0.00043*	*0.00010*	*0.00005*	*0.00003*		*0.00002*	*0.00002*	*0.00003*	*0.00003*
(12) Middle School	-0.070	-0.035	0.003	0.042	0.073	-0.088	-0.034	0.015	0.177	-0.197	-0.185	1	*0.00010*	*0.00050*	*0.00010*	*0.00007*	*0.00005*		*0.00007*	*0.00007*	*0.00003*	*0.00007*
(13) High School	-0.047	-0.032	0.088	0.084	0.023	-0.097	-0.071	0.049	0.100	-0.147	-0.137	-0.101	1	*0.00014*	*0.00003*	*0.00002*	*0.00001*		*0.00005*	*0.00002*	*0.00003*	*0.00002*
(14) Basic Vocational School	0.095	0.076	-0.076	-0.021	-0.018	0.112	-0.022	-0.064	0.006	-0.287	-0.269	-0.198	-0.147	1	*0.00017*	*0.00011*	*0.00008*		*0.00003*	*0.00001*	*0.00002*	*0.00002*
(15) Advanced Vocational School	-0.032	0.015	0.025	0.065	0.036	-0.060	-0.059	0.012	-0.010	-0.151	-0.141	-0.104	-0.077	-0.151	1	*0.00004*	*0.00001*		*0.00003*	*0.00009*	*0.00008*	*0.00007*
(16) Technical College or University	-0.064	-0.013	0.111	0.136	-0.025	-0.098	-0.083	0.046	0.030	-0.126	-0.117	-0.086	-0.064	-0.126	-0.066	1	*0.00001*		*0.00006*	*0.00004*	*0.00004*	*0.00004*
(17) Graduate School	0.043	-0.026	0.291	0.021	-0.064	-0.083	-0.073	0.089	-0.060	-0.103	-0.097	-0.071	-0.053	-0.104	-0.054	-0.045	1		*0.00004*	*0.00001*	*0.00001*	*0.00001*
(18) Paid Days	0.008	0.239	0.054	0.077	-0.048	0.047	-0.110	-0.045	-0.297	-0.095	0.094	-0.072	-0.041	0.066	0.026	-0.001	0.010	1	*0.00000*	*0.00000*	*0.00000*	*0.00002*
(19) Log Real Annual Compensation	0.191	0.529	0.207	0.096	-0.120	0.055	-0.168	0.098	-0.165	-0.116	-0.013	-0.026	0.012	0.048	0.049	0.048	0.082	0.110	1	*0.00000*	*0.00000*	*0.00001*
(20) Log Sales	0.060	0.013	0.044	0.068	0.016	-0.042	-0.063	0.030	-0.007	-0.054	-0.038	0.032	0.030	-0.003	0.032	0.025	0.041	0.067	0.131	1	*0.00001*	*0.00001*
(21) Log Capital Stock	0.066	0.126	0.046	0.077	0.006	-0.024	-0.082	0.004	-0.072	-0.074	-0.017	0.018	0.025	0.011	0.031	0.024	0.044	0.157	0.185	0.894	1	*0.00001*
(22) Log Average Employment	0.063	-0.033	-0.004	0.039	-0.019	-0.016	0.003	0.002	-0.059	-0.036	-0.011	0.011	0.008	0.003	0.021	0.010	0.027	0.042	0.073	0.929	0.848	1

TABLE 5 (continued)
CORRELATIONS IN MASKED DATA (Between Implicate Variance Above Diagonal)

									MASKED VARIABLES													
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
(1) Male	1								*0.00003*	*0.02205*	*0.00458*	*0.00237*	*0.00077*	*0.00748*	*0.00363*	*0.00161*	*0.00014*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(2) Full Time Employee	0.167	1							*0.00000*	*0.00069*	*0.00020*	*0.00010*	*0.00001*	*0.00016*	*0.00006*	*0.00002*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(3) Engineer, Professional, or Manager	0.097	-0.008	1						*0.00000*	*0.00014*	*0.00006*	*0.00010*	*0.00010*	*0.00008*	*0.00017*	*0.00017*	*0.00013*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(4) Technician or Technical White Collar	0.001	0.031	-0.155	1					*0.00000*	*0.00011*	*0.00007*	*0.00005*	*0.00001*	*0.00013*	*0.00012*	*0.00008*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(5) Other White Collar	-0.380	-0.088	-0.205	-0.309	1				*0.00000*	*0.00325*	*0.00064*	*0.00042*	*0.00012*	*0.00160*	*0.00085*	*0.00016*	*0.00001*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(6) Skilled Blue Collar	0.314	0.132	-0.176	-0.265	-0.351	1			*0.00000*	*0.00321*	*0.00089*	*0.00020*	*0.00004*	*0.00147*	*0.00022*	*0.00010*	*0.00001*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(7) Unskilled Blue Collar	0.030	-0.063	-0.158	-0.238	-0.316	-0.271	1		*0.00000*	*0.00022*	*0.00014*	*0.00008*	*0.00002*	*0.00016*	*0.00004*	*0.00004*	*0.00001*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(8) Works in Ile-de-France	-0.018	-0.007	0.133	0.066	0.034	-0.092	-0.104	1	*0.00000*	*0.00022*	*0.00002*	*0.00003*	*0.00005*	*0.00004*	*0.00003*	*0.00009*	*0.00006*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(9) Year of Birth	-0.075	-0.110	-0.091	-0.050	0.086	-0.057	0.079	-0.035	1	*0.00096*	*0.00013*	*0.00065*	*0.00015*	*0.00007*	*0.00006*	*0.00011*	*0.00002*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(10) No Diploma	0.079	-0.037	-0.099	-0.116	-0.040	0.086	0.141	-0.022	0.073	1	*0.00154*	*0.00040*	*0.00043*	*0.00189*	*0.00061*	*0.00039*	*0.00020*	*0.00030*	*0.00072*	*0.00000*	*0.00000*	*0.00001*
(11) Elementary School	-0.013	0.027	-0.078	-0.058	-0.001	0.071	0.041	-0.034	-0.251	-0.264	1	*0.00076*	*0.00015*	*0.00045*	*0.00038*	*0.00023*	*0.00007*	*0.00003*	*0.00023*	*0.00000*	*0.00000*	*0.00000*
(12) Middle School	-0.089	-0.039	0.001	0.037	0.071	-0.087	-0.026	0.015	0.158	-0.197	-0.180	1	*0.00036*	*0.00104*	*0.00028*	*0.00026*	*0.00015*	*0.00026*	*0.00014*	*0.00000*	*0.00000*	*0.00000*
(13) High School	-0.056	-0.031	0.079	0.079	0.026	-0.095	-0.065	0.052	0.091	-0.149	-0.135	-0.102	1	*0.00024*	*0.00007*	*0.00005*	*0.00002*	*0.00006*	*0.00004*	*0.00000*	*0.00000*	*0.00000*
(14) Basic Vocational School	0.063	0.069	-0.072	-0.023	-0.010	0.093	-0.013	-0.060	-0.002	-0.287	-0.258	-0.196	-0.147	1	*0.00024*	*0.00018*	*0.00008*	*0.00009*	*0.00021*	*0.00000*	*0.00000*	*0.00000*
(15) Advanced Vocational School	-0.003	0.019	0.027	0.061	0.021	-0.050	-0.052	0.006	-0.014	-0.158	-0.143	-0.108	-0.081	-0.155	1	*0.00009*	*0.00003*	*0.00001*	*0.00017*	*0.00000*	*0.00000*	*0.00000*
(16) Technical College or University	-0.087	-0.015	0.097	0.124	-0.011	-0.096	-0.078	0.044	0.031	-0.126	-0.114	-0.086	-0.064	-0.123	-0.068	1	*0.00003*	*0.00005*	*0.00008*	*0.00001*	*0.00000*	*0.00001*
(17) Graduate School	0.046	-0.029	0.278	0.023	-0.061	-0.080	-0.071	0.095	-0.050	-0.106	-0.095	-0.072	-0.054	-0.104	-0.057	-0.046	1	*0.00000*	*0.00003*	*0.00000*	*0.00000*	*0.00000*
(18) Paid Days	0.011	0.246	0.055	0.078	-0.051	0.050	-0.111	-0.047	-0.299	-0.092	0.094	-0.071	-0.038	0.067	0.027	-0.003	0.003	1	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(19) Log Real Annual Compensation	0.193	0.535	0.210	0.098	-0.121	0.056	-0.171	0.100	-0.164	-0.102	-0.009	-0.030	0.009	0.042	0.050	0.037	0.077	0.113	1	*0.00000*	*0.00000*	*0.00000*
(20) Log Sales	0.008	-0.005	0.003	-0.006	-0.007	0.006	0.005	0.004	0.013	0.003	-0.005	0.002	0.001	-0.001	-0.001	-0.003	0.002	-0.010	-0.004	1	*0.00000*	*0.00001*
(21) Log Capital Stock	0.008	-0.005	0.008	-0.005	-0.005	0.002	0.001	-0.003	0.017	0.004	-0.007	0.004	0.002	-0.002	0.000	-0.003	0.003	-0.008	-0.002	0.737	1	*0.00000*
(22) Log Average Employment	0.009	-0.001	0.008	-0.003	-0.008	0.005	0.001	0.006	0.018	0.003	-0.008	0.003	0.003	-0.001	0.000	-0.002	0.002	-0.007	0.001	0.773	0.661	1

TABLE 5 (continued)
CORRELATIONS IN SIMULATED DATA (Between Implicate Variance Above Diagonal)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
									SIMULATED VARIABLES													
(1) Male	1								*0.00052*	*0.03068*	*0.00383*	*0.00423*	*0.00416*	*0.00715*	*0.00402*	*0.00326*	*0.00056*	*0.00006*	*0.00036*	*0.00000*	*0.00000*	*0.00000*
(2) Full Time Employee	0.167	1							*0.00001*	*0.00071*	*0.00009*	*0.00008*	*0.00008*	*0.00025*	*0.00012*	*0.00009*	*0.00001*	*0.00014*	*0.00064*	*0.00000*	*0.00000*	*0.00000*
(3) Engineer, Professional, or Manager	0.097	-0.008	1						*0.00002*	*0.00029*	*0.00007*	*0.00014*	*0.00042*	*0.00024*	*0.00020*	*0.00055*	*0.00042*	*0.00000*	*0.00001*	*0.00000*	*0.00000*	*0.00000*
(4) Technician or Technical White Collar	0.001	0.031	-0.155	1					*0.00002*	*0.00015*	*0.00009*	*0.00006*	*0.00013*	*0.00021*	*0.00021*	*0.00003*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(5) Other White Collar	-0.380	-0.088	-0.205	-0.309	1				*0.00008*	*0.00391*	*0.00045*	*0.00059*	*0.00056*	*0.00139*	*0.00076*	*0.00032*	*0.00006*	*0.00001*	*0.00005*	*0.00000*	*0.00000*	*0.00000*
(6) Skilled Blue Collar	0.314	0.132	-0.176	-0.265	-0.351	1			*0.00011*	*0.00439*	*0.00070*	*0.00022*	*0.00009*	*0.00232*	*0.00031*	*0.00010*	*0.00002*	*0.00001*	*0.00005*	*0.00000*	*0.00000*	*0.00000*
(7) Unskilled Blue Collar	0.030	-0.063	-0.158	-0.238	-0.316	-0.271	1		*0.00002*	*0.00019*	*0.00018*	*0.00007*	*0.00003*	*0.00019*	*0.00005*	*0.00003*	*0.00001*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(8) Works in Ile-de-France	-0.018	-0.007	0.133	0.066	0.034	-0.092	-0.104	1	*0.00001*	*0.00006*	*0.00001*	*0.00007*	*0.00006*	*0.00003*	*0.00002*	*0.00007*	*0.00006*	*0.00000*	*0.00000*	*0.00000*	*0.00000*	*0.00000*
(9) Year of Birth	-0.094	-0.039	-0.099	-0.063	0.079	-0.034	0.081	-0.058	1	*0.00065*	*0.00007*	*0.00007*	*0.00023*	*0.00015*	*0.00016*	*0.00019*	*0.00004*	*0.00001*	*0.00002*	*0.00000*	*0.00000*	*0.00001*
(10) No Diploma	0.083	0.023	-0.102	-0.126	-0.049	0.120	0.127	-0.013	0.018	1	*0.00158*	*0.00096*	*0.00061*	*0.00047*	*0.00152*	*0.00080*	*0.00018*	*0.00004*	*0.00122*	*0.00001*	*0.00000*	*0.00001*
(11) Elementary School	-0.035	-0.004	-0.075	-0.063	0.030	0.033	0.048	-0.035	0.012	-0.266	1	*0.00038*	*0.00018*	*0.00112*	*0.00020*	*0.00011*	*0.00007*	*0.00001*	*0.00014*	*0.00000*	*0.00000*	*0.00001*
(12) Middle School	-0.087	-0.019	0.027	0.062	0.055	-0.089	-0.049	0.020	0.005	-0.255	-0.164	1	*0.00037*	*0.00165*	*0.00013*	*0.00015*	*0.00010*	*0.00000*	*0.00023*	*0.00000*	*0.00001*	*0.00000*
(13) High School	-0.062	-0.023	0.101	0.102	0.011	-0.102	-0.078	0.046	-0.018	-0.183	-0.117	-0.114	1	*0.00074*	*0.00009*	*0.00004*	*0.00005*	*0.00001*	*0.00024*	*0.00000*	*0.00001*	*0.00000*
(14) Basic Vocational School	0.079	0.038	-0.072	-0.028	-0.011	0.087	0.001	-0.058	0.008	-0.316	-0.205	-0.199	-0.142	1	*0.00032*	*0.00029*	*0.00030*	*0.00001*	*0.00020*	*0.00000*	*0.00000*	*0.00000*
(15) Advanced Vocational School	-0.025	-0.011	0.009	0.050	0.037	-0.051	-0.044	-0.001	0.000	-0.186	-0.118	-0.113	-0.081	-0.141	1	*0.00004*	*0.00003*	*0.00001*	*0.00026*	*0.00000*	*0.00001*	*0.00000*
(16) Technical College or University	-0.061	-0.016	0.104	0.122	-0.024	-0.090	-0.073	0.039	-0.023	-0.137	-0.087	-0.084	-0.060	-0.105	-0.060	1	*0.00001*	*0.00002*	*0.00031*	*0.00000*	*0.00000*	*0.00000*
(17) Graduate School	0.024	-0.034	0.244	0.026	-0.053	-0.075	-0.063	0.079	-0.046	-0.114	-0.073	-0.071	-0.051	-0.089	-0.051	-0.037	1	*0.00001*	*0.00009*	*0.00000*	*0.00000*	*0.00000*
(18) Paid Days	0.013	0.258	0.039	0.068	-0.053	0.053	-0.092	-0.040	-0.049	-0.014	-0.006	0.002	0.003	0.012	0.003	0.007	-0.001	1	*0.00007*	*0.00000*	*0.00000*	*0.00000*
(19) Log Real Annual Compensation	0.198	0.549	0.218	0.101	-0.124	0.056	-0.177	0.085	-0.069	-0.027	-0.034	0.001	0.021	0.010	0.005	0.025	0.041	0.199	1	*0.00000*	*0.00000*	*0.00000*
(20) Log Sales	-0.005	-0.003	0.003	0.004	0.004	-0.008	-0.006	0.011	0.000	-0.001	0.000	0.001	0.001	-0.002	0.000	0.001	0.002	-0.001	-0.002	1	*0.00014*	*0.00041*
(21) Log Capital Stock	-0.005	0.006	0.002	0.004	0.003	-0.003	-0.007	0.012	-0.001	-0.001	0.000	0.000	0.001	-0.001	0.000	0.002	0.002	0.001	0.004	0.562	1	*0.00061*
(22) Log Average Employment	0.010	0.017	0.004	-0.006	-0.018	0.016	0.011	0.019	-0.001	0.004	-0.002	-0.002	-0.001	0.001	-0.001	-0.001	0.001	0.005	0.014	0.358	0.382	1

Notes: N=3,213,374
Sources: Authors' calculations based on the INSEE DADS and EDP data.

TABLE 6: BIVARIATE STATISTICS IN THE FIRM DATA

CORRELATIONS IN COMPLETED DATA
(Between Implicate Variance Above Diagonal)

	(1)	(2)	(3)
(1) Log Sales	1	*0.000001*	*0.000013*
(2) Log Capital Stock	0.733	1	*0.000003*
(3) Log Average Employment	0.767	0.656	1

CORRELATIONS IN MASKED DATA
(Between Implicate Variance Above Diagonal)

	(1)	(2)	(3)
(1) Log Sales	1	*0.000001*	*0.000011*
(2) Log Capital Stock	0.737	1	*0.000004*
(3) Log Average Employment	0.773	0.661	1

CORRELATIONS IN SIMULATED DATA
(Between Implicate Variance Above Diagonal)

	(1)	(2)	(3)
(1) Log Sales	1	*0.000138*	*0.000410*
(2) Log Capital Stock	0.562	1	*0.000614*
(3) Log Average Employment	0.358	0.382	1

Notes: N=470,812 in completed and masked data; N=470,378 in the simulated data.
Sources: Authors' calculations based on the INSEE EAE data.

TABLE 7: SUMMARY OF COEFFICIENT ESTIMATES FOR THE ANALYSIS OF A LINEAR MODEL PREDICTING LOG REAL ANNUAL WAGE RATES WITH FIXED PERSON AND FIRM EFFECTS

	COMPLETED DATA					MASKED DATA					SIMULATED DATA				
TIME-VARYING OBSERVABLES	Average Coefficient	Average Variance of Coefficient	Between-Implicate Variance of Coefficient	Total Variance of Coefficient	Standard Error of Average Coefficient	Average Coefficient	Average Variance of Coefficient	Between-Implicate Variance of Coefficient	Total Variance of Coefficient	Standard Error of Average Coefficient	Average Coefficient	Average Variance of Coefficient	Between-Implicate Variance of Coefficient	Total Variance of Coefficient	Standard Error of Average Coefficient
Male x Experience	0.0757	6.88E-07	3.86E-06	4.93E-06	0.0022	0.0470	5.93E-07	4.19E-07	1.05E-06	0.0010	0.0128	6.85E-07	2.31E-05	2.61E-05	0.0051
Male x (Experience^2)/100	-0.2676	4.24E-05	1.54E-04	2.12E-04	0.0146	-0.0858	4.27E-05	4.47E-05	9.19E-05	0.0096	-0.0257	5.63E-05	8.29E-05	1.47E-04	0.0121
Male x (Experience^3)/1000	0.0519	4.17E-06	1.06E-05	1.58E-05	0.0040	0.0055	4.76E-06	7.63E-06	1.32E-05	0.0036	0.0066	6.62E-06	6.09E-06	1.33E-05	0.0036
Male x (Experience^4)/10000	-0.0041	4.76E-08	8.96E-08	1.46E-07	0.0004	0.0001	6.02E-08	1.28E-07	2.01E-07	0.0004	-0.0007	8.43E-08	7.51E-08	1.67E-07	0.0004
Female x Experience	0.0386	1.46E-06	8.82E-07	2.44E-06	0.0016	0.0329	1.31E-06	5.36E-06	7.20E-06	0.0027	0.0167	1.64E-06	7.36E-06	9.74E-06	0.0031
Female x (Experience^2)/100	-0.0709	9.66E-05	4.39E-05	1.45E-04	0.0120	-0.0405	1.01E-04	4.20E-04	5.63E-04	0.0237	-0.0425	1.50E-04	2.69E-04	4.45E-04	0.0211
Female x (Experience^3)/1000	0.0075	9.91E-06	3.95E-06	1.42E-05	0.0038	0.0006	1.18E-05	4.48E-05	6.11E-05	0.0078	0.0121	1.92E-05	3.17E-05	5.41E-05	0.0074
Female x (Experience^4)/10000	-0.0002	1.16E-07	4.08E-08	1.61E-07	0.0004	0.0004	1.53E-07	5.27E-07	7.33E-07	0.0009	-0.0012	2.64E-07	3.80E-07	6.82E-07	0.0008
Male x Works in Ile-de-France	0.0465	1.29E-05	5.12E-07	1.35E-05	0.0037	0.0512	1.32E-05	1.67E-06	1.51E-05	0.0039	0.0690	2.45E-05	1.17E-05	3.74E-05	0.0061
Female x Works in Ile-de-France	0.0391	4.52E-05	8.91E-07	4.62E-05	0.0068	0.0430	4.67E-05	9.50E-06	5.71E-05	0.0076	0.0840	8.76E-05	5.97E-05	1.53E-04	0.0124
Log Sales	0.0024	1.21E-05	3.02E-05	4.54E-05	0.0067	0.0100	1.25E-05	5.17E-05	6.93E-05	0.0083	0.1337	1.40E-04	5.43E-03	6.11E-03	0.0782
(Log Sales)^2	0.0005	1.86E-08	6.72E-08	9.25E-08	0.0003	0.0004	1.89E-08	1.15E-07	1.45E-07	0.0004	-0.0051	2.96E-07	8.53E-06	9.68E-06	0.0031
TIME-INVARIANT OBSERVABLES															
Male	-0.067	2.41E-05	1.79E-04	2.21E-04	0.0149	0.046	3.41E-05	5.22E-04	6.08E-04	0.0247	0.244	2.53E-05	6.69E-03	7.39E-03	0.0859
Male x Elementary School	0.002	1.90E-05	1.54E-04	1.88E-04	0.0137	0.017	2.24E-05	2.98E-04	3.50E-04	0.0187	0.016	2.56E-05	6.32E-05	9.52E-05	0.0098
Male x Middle School	0.179	2.72E-05	7.24E-04	8.24E-04	0.0287	0.155	3.49E-05	4.99E-04	5.84E-04	0.0242	0.157	3.01E-05	5.27E-04	6.10E-04	0.0247
Male x High School	0.315	4.31E-05	8.10E-04	9.34E-04	0.0306	0.316	5.26E-05	3.12E-04	3.96E-04	0.0199	0.274	5.30E-05	5.31E-04	6.37E-04	0.0252
Male x Basic Vocational School	0.149	1.68E-05	2.85E-04	3.30E-04	0.0182	0.130	1.96E-05	2.16E-04	2.57E-04	0.0160	0.051	1.83E-05	1.21E-04	1.51E-04	0.0123
Male x Advanced Vocational School	0.291	3.92E-05	2.19E-04	2.80E-04	0.0167	0.268	4.09E-05	9.53E-04	1.09E-03	0.0330	0.162	4.09E-05	3.48E-04	4.23E-04	0.0206
Male x Technical College or University	0.490	6.97E-05	6.33E-04	7.66E-04	0.0277	0.483	9.26E-05	4.79E-04	6.20E-04	0.0249	0.382	1.04E-04	6.21E-04	7.87E-04	0.0281
Male x Graduate School	0.760	7.02E-05	2.57E-04	3.53E-04	0.0188	0.716	7.89E-05	8.85E-04	1.05E-03	0.0324	0.511	1.01E-04	6.37E-04	8.01E-04	0.0283
Female x Elementary School	-0.108	3.55E-05	1.18E-04	1.65E-04	0.0129	-0.061	4.89E-05	2.66E-04	3.41E-04	0.0185	0.008	4.72E-05	8.52E-05	1.41E-04	0.0119
Female x Middle School	0.142	3.89E-05	2.29E-04	2.91E-04	0.0171	0.127	5.04E-05	2.54E-04	3.30E-04	0.0182	0.082	4.08E-05	1.00E-04	1.51E-04	0.0123
Female x High School	0.233	6.29E-05	6.11E-04	7.35E-04	0.0271	0.260	8.05E-05	2.12E-04	3.14E-04	0.0177	0.169	7.51E-05	1.55E-04	2.45E-04	0.0157
Female x Basic Vocational School	0.110	3.93E-05	1.40E-04	1.93E-04	0.0139	0.109	5.14E-05	2.61E-04	3.38E-04	0.0184	0.058	4.93E-05	6.10E-05	1.16E-04	0.0108
Female x Advanced Vocational School	0.187	6.65E-05	2.88E-04	3.83E-04	0.0196	0.199	9.28E-05	4.03E-04	5.36E-04	0.0232	0.104	7.99E-05	1.23E-04	2.15E-04	0.0147
Female x Technical College or University	0.364	8.89E-05	1.72E-04	2.78E-04	0.0167	0.361	1.05E-04	5.26E-04	6.83E-04	0.0261	0.226	1.21E-04	1.94E-04	3.34E-04	0.0183
Female x Graduate School	0.460	2.23E-04	2.52E-04	5.00E-04	0.0224	0.484	2.85E-04	8.73E-04	1.25E-03	0.0353	0.351	2.72E-04	9.32E-04	1.30E-03	0.0360

Notes: The model is a linear analysis of covariance with covariates listed in the table, unrestricted fixed person and firm effects. The dependent variable is the logarithm of annual full time, full year compensation.
Sources: Authors' calculations based on INSEE DADS, EDP, and EAE data.

TABLE 8: CORRELATION OF ESTIMATED EFFECTS FROM THE LOG WAGE RATE REGRESSION MODEL

	Log wage	Time-varying Observables	Person Effect	Time-invariant Observables	Rest of Person Effect	Firm Effect	Residual
AVERAGE CORRELATIONS IN COMPLETED DATA (Between-Implicate Variance Above Diagonal)							
Log Real Annual Compensation	1	*0.000021*	*0.000023*	*0.000079*	*0.000042*	*0.000006*	*0.000000*
Time-Varying Obsevables	0.417	1	*0.000017*	*0.000110*	*0.000022*	*0.000004*	*0.000000*
Person Effect	0.517	-0.140	1	*0.000087*	*0.000017*	*0.000069*	*0.000000*
Time-Invariant Observables	0.336	-0.146	0.383	1	*0.000005*	*0.000011*	*0.000000*
Rest of Person Effect	0.430	-0.095	0.934	0.029	1	*0.000096*	*0.000000*
Firm Effect	0.258	0.042	-0.466	0.105	-0.545	1	*0.000000*
Residual	0.393	0.000	0.000	0.000	0.000	0.000	1
AVERAGE CORRELATIONS IN MASKED DATA (Between-Implicate Variance Above Diagonal)							
Log Real Annual Compensation	1	*0.000067*	*0.000608*	*0.000083*	*0.000344*	*0.000036*	*0.000001*
Time-Varying Obsevables	0.351	1	*0.000060*	*0.000539*	*0.000029*	*0.000048*	*0.000000*
Person Effect	0.503	-0.172	1	*0.000411*	*0.000062*	*0.001582*	*0.000000*
Time-Invariant Observables	0.359	-0.086	0.339	1	*0.000015*	*0.000049*	*0.000000*
Rest of Person Effect	0.409	-0.153	0.944	0.011	1	*0.001336*	*0.000000*
Firm Effect	0.258	0.044	-0.510	0.105	-0.579	1	*0.000000*
Residual	0.402	0.000	0.000	0.000	0.000	0.000	1
AVERAGE CORRELATIONS IN SIMULATED DATA (Between-Implicate Variance Above Diagonal)							
Log Real Annual Compensation	1	*0.008058*	*0.003133*	*0.000524*	*0.001743*	*0.000161*	*0.000032*
Time-Varying Obsevables	0.109	1	*0.003557*	*0.049421*	*0.002011*	*0.000176*	*0.000000*
Person Effect	0.479	-0.173	1	*0.004107*	*0.000425*	*0.004680*	*0.000000*
Time-Invariant Observables	0.399	-0.153	0.380	1	*0.000154*	*0.000130*	*0.000000*
Rest of Person Effect	0.369	-0.120	0.936	0.035	1	*0.003206*	*0.000000*
Firm Effect	0.202	0.030	-0.598	0.065	-0.670	1	*0.000000*
Residual	0.573	0.000	0.000	0.000	0.000	0.000	1

Notes: Based on statistics produced by the model summarized in Table 7.
Sources: Authors' calculations based on INSEE DADS, EDP and EAE data.

TABLE 9: SUMMARY OF COEFFICIENT ESTIMATES FOR THE ANALYSIS OF A MODEL PREDICTING FULL TIME EMPLOYMENT

	COMPLETED DATA					MASKED DATA					SIMULATED DATA				
PREDICTOR VARIABLE	Average Coefficient	Average Variance	Between Implicate Variance	Total Variance	Standard Error of Average Coefficient	Average Coefficient	Average Variance	Between Implicate Variance	Total Variance	Standard Error of Average Coefficient	Average Coefficient	Average Variance	Between Implicate Variance	Total Variance	Standard Error of Average Coefficient
Intercept	-6.977	0.00192	0.05725	0.06489	0.255	-7.418	0.00218	0.05747	0.06539	0.256	-10.550	0.00349	0.24928	0.27771	0.527
Male	2.216	0.00251	0.00251	0.00528	0.073	1.978	0.00292	0.00828	0.01202	0.110	0.978	0.00269	0.39920	0.44181	0.665
Male x Experience	-0.001	1.6755E-07	6.2557E-07	8.5567E-07	0.001	-0.002	1.9644E-07	1.3467E-06	1.6778E-06	0.001	-0.003	1.7804E-07	6.6510E-07	9.0965E-07	0.001
Male x Works in Ile-de-France	-0.203	0.00011	0.00001	0.00011	0.011	-0.220	0.00011	0.00005	0.00017	0.013	-0.183	0.00011	0.00017	0.00030	0.017
Male x Elementary School	0.327	0.00019	0.00065	0.00090	0.030	0.323	0.00021	0.00128	0.00162	0.040	0.011	0.00023	0.00434	0.00500	0.071
Male x Middle School	-0.236	0.00024	0.00163	0.00203	0.045	-0.202	0.00029	0.00170	0.00216	0.046	-0.101	0.00029	0.00284	0.00342	0.058
Male x High School	-0.269	0.00041	0.00201	0.00262	0.051	-0.251	0.00046	0.00199	0.00266	0.052	-0.201	0.00056	0.00249	0.00330	0.057
Male x Basic Vocational School	0.253	0.00016	0.00086	0.00111	0.033	0.280	0.00017	0.00197	0.00234	0.048	0.115	0.00017	0.00292	0.00339	0.058
Male x Advanced Vocational School	-0.074	0.00038	0.00112	0.00161	0.040	-0.011	0.00039	0.00157	0.00211	0.046	-0.114	0.00039	0.00177	0.00233	0.048
Male x Technical College or University	-0.146	0.00075	0.00107	0.00193	0.044	-0.122	0.00096	0.00255	0.00377	0.061	-0.083	0.00121	0.00288	0.00438	0.066
Male x Graduate School	-0.403	0.00085	0.00220	0.00328	0.057	-0.447	0.00084	0.00284	0.00396	0.063	-0.375	0.00104	0.00311	0.00446	0.067
Male x Engineer, Professional or Manager	-0.565	0.00050	0.00036	0.00090	0.030	-0.706	0.00053	0.00026	0.00082	0.029	-0.974	0.00053	0.00990	0.01142	0.107
Male x Technical or Technical White Collar	0.362	0.00030	0.00004	0.00034	0.018	0.294	0.00032	0.00011	0.00044	0.021	0.237	0.00032	0.00312	0.00375	0.061
Male x Skilled Blue Collar	0.389	0.00020	0.00025	0.00047	0.022	0.370	0.00021	0.00038	0.00063	0.025	0.504	0.00022	0.00131	0.00166	0.041
Male x Unskilled Blue Collar	0.091	0.00020	0.00034	0.00057	0.024	0.099	0.00022	0.00036	0.00061	0.025	0.090	0.00022	0.00029	0.00054	0.023
Male x Log Real Annual Compensation	1.949	0.00006	0.00005	0.00012	0.011	2.129	0.00007	0.00017	0.00025	0.016	2.180	0.00006	0.00776	0.00860	0.093
Female x Experience	-0.012	2.0232E-07	3.2116E-07	5.5560E-07	0.001	-0.011	2.3378E-07	4.6906E-07	7.4974E-07	0.001	-0.009	2.2073E-07	4.4407E-07	7.0920E-07	0.001
Female x Works in Ile-de-France	-0.046	0.00012	0.00001	0.00013	0.011	-0.065	0.00013	0.00004	0.00018	0.013	0.063	0.00013	0.00008	0.00022	0.015
Female x Elementary School	0.368	0.00021	0.00065	0.00093	0.030	0.352	0.00027	0.00276	0.00331	0.058	-0.002	0.00026	0.00333	0.00392	0.063
Female x Middle School	-0.038	0.00024	0.00149	0.00189	0.043	-0.006	0.00029	0.00244	0.00298	0.055	-0.032	0.00024	0.00265	0.00315	0.056
Female x High School	-0.128	0.00042	0.00138	0.00194	0.044	-0.121	0.00047	0.00215	0.00283	0.053	-0.084	0.00046	0.00363	0.00445	0.067
Female x Basic Vocational School	0.239	0.00023	0.00058	0.00087	0.029	0.261	0.00028	0.00196	0.00243	0.049	-0.005	0.00027	0.00204	0.00252	0.050
Female x Advanced Vocational School	0.040	0.00040	0.00105	0.00156	0.039	0.085	0.00051	0.00409	0.00500	0.071	-0.122	0.00044	0.00359	0.00438	0.066
Female x Technical College or University	-0.173	0.00063	0.00223	0.00309	0.056	-0.090	0.00067	0.00496	0.00613	0.078	-0.082	0.00080	0.00600	0.00740	0.086
Female x Graduate School	-0.566	0.00151	0.00137	0.00302	0.055	-0.400	0.00167	0.00871	0.01125	0.106	-0.331	0.00164	0.00762	0.01002	0.100
Female x Engineer, Professional or Manager	-0.758	0.00077	0.00019	0.00098	0.031	-0.882	0.00080	0.00074	0.00162	0.040	-0.905	0.00081	0.01085	0.01274	0.113
Female x Technical or Technical White Collar	0.182	0.00024	0.00022	0.00048	0.022	0.135	0.00025	0.00030	0.00059	0.024	0.086	0.00024	0.00053	0.00082	0.029
Female x Skilled Blue Collar	1.116	0.00034	0.00055	0.00094	0.031	1.112	0.00036	0.00058	0.00100	0.032	1.172	0.00036	0.00099	0.00145	0.038
Female x Unskilled Blue Collar	0.625	0.00014	0.00066	0.00087	0.030	0.643	0.00015	0.00073	0.00095	0.031	0.590	0.00015	0.00213	0.00250	0.050
Female x Log Real Annual Compensation	2.385	0.00010	0.00005	0.00015	0.012	2.504	0.00011	0.00026	0.00039	0.020	2.328	0.00009	0.02349	0.02592	0.161
Log Sales	-0.245	0.00002	0.00190	0.00210	0.046	-0.240	0.00002	0.00188	0.00209	0.046	0.082	0.00003	0.00051	0.00059	0.024
Log Capital Stock	0.338	0.00000	0.00004	0.00004	0.007	0.330	0.00000	0.00005	0.00006	0.008	0.182	0.00001	0.00043	0.00049	0.022
Log Average Employment	-0.285	0.00001	0.00287	0.00317	0.056	-0.282	0.00002	0.00294	0.00325	0.057	0.038	0.00004	0.00132	0.00149	0.039

Notes: The model was estimated with maximum likelihood logistic regression. The dependent variable is full time employment status (not full time is the reference category).
Sources: Authors' calculations based on the INSEE DADS, EDP and EAE data.