

Proceedings of the  
34th International Workshop  
on Statistical Modelling  
Volume I

July 7-12, 2019  
Guimarães, Portugal

Title: Proceedings of the 34th International Workshop on Statistical Modelling  
Volume I,  
Guimarães, July 7-12, 2019,  
Luís Meira-Machado, Gustavo Soutinho (editors),  
Guimarães, 2019.

## **Editors:**

Luís Meira-Machado, [lmachado@math.uminho.pt](mailto:lmachado@math.uminho.pt)  
University of Minho  
Dep. Mathematics and Applications  
4810-058 Azurém - Guimarães  
Portugal

Gustavo Soutinho, [gustavo.soutinho@ispup.up.pt](mailto:gustavo.soutinho@ispup.up.pt)  
EPIUnit, ICBADS, University of Porto  
Rua das Taipas 135  
4050-600 Porto  
Portugal

ISBN 978-989-20-9528-8  
Printed by Instituto Nacional de Estatística

**Part III - Special Session Devoted to  
Statistics Portugal**



# Synthetic data as Public Use Files: an application to the Household Budget Survey

Inês Rodrigues<sup>1</sup>

<sup>1</sup> Instituto Nacional de Estatística - Statistics Portugal, Lisboa, Portugal

E-mail for correspondence: `ines.rodrigues@ine.pt`

**Abstract:** A methodology for producing Public Use Files (PUF) for the Household Budget Survey by generating synthetic data is presented. Parametric (multinomial logistic and log-linear regressions) and non-parametric methods (classification and regression trees) were used for generating the main identifying variables, as well as income and expenditure totals. The two approaches were compared with a focus on the risk of disclosing confidential information from the PUF.

**Keywords:** PUF; HBS; Confidentiality; Synthetic data.

## 1 Introduction

Public Use Files (PUF) include data on individual statistical units and are prepared to be of public access. PUF are intended to be used for education or test purposes - e.g., by researchers when developing their application to access microdata files for research use, the so-called SUF (Scientific Use Files). The aim of this work is to compare a parametric and a non-parametric approach for producing PUF for the Household Budget Survey based on synthetic data, namely regarding the resulting disclosure risk.

## 2 Producing PUF by generating synthetic data

### 2.1 A parametric and a non-parametric approach

Let  $D$  denote an original microdata set, including a set of  $k$  variables represented by  $Y$ , whose relationships are intended to be preserved. Following Raghunathan et al. (2001), the joint conditional density of  $Y_1, Y_2, \dots, Y_k$

---

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

given a set of background variables  $X$ , can be factored as:

$$f(Y_1, \dots, Y_k | X, \theta_1, \dots, \theta_k) = f_1(Y_1 | X, \theta_1) \prod_{v=2}^k f_v(Y_v | X, Y_1, \dots, Y_{v-1}, \theta_v)$$

where  $f_v$ ,  $v = 1, \dots, k$  are the conditional density functions and  $\theta_v$  is a vector of parameters in the conditional distribution (e.g., regression coefficients). Each conditional distribution is modeled by a given appropriate regression model (e.g., linear, logistic or log-linear regression, if  $Y_v$  is a continuous, binary or count variable, respectively).

On the other hand, the CART (classification and regression trees) algorithm provides good results as a non-parametric approach to generate synthetic data (Dreschler and Reiter (2011)). As described by Nowok et al. (2017), it is based on the recursive partition of the original dataset into groups with increasingly homogeneous outcome. Splits are defined based on *yes/no* questions concerning the predictors. In each final group (leaf), values approximate the conditional distribution of the predicted variable for units with predictors meeting the criteria that define that group. Synthetic values are generated by sampling from the appropriate leaf. CART can be used to simulate each variable sequentially, by conditioning on already generated variables, as in the parametric approach.

## 2.2 Disclosure risk

Following Loong et al. (2013), we assume the user knows which units are included in  $D$  and their values regarding  $m$  indirect identifiers. Based on this  $m$  variables, the user attempts to obtain information on a confidential variable,  $T$ , from the synthetic microdata set  $D'$ . Let  $n_R$  and  $n_S$  denote the number of units in  $D$  and  $D'$ , respectively. We denote by  $w_{iq}$  the value of the indirect identifier  $q$  for unit  $i$  in  $D$  ( $i = 1, \dots, n_R$ ;  $q = 1, \dots, m$ ) and by  $w_{jq}$  the value of the same identifier for unit  $j$  in  $D'$ . Let  $R_{ij} = 1$  ( $i = 1, \dots, n_R$ ;  $j = 1, \dots, n_S$ ) if  $w_{iq} = w_{jq} \forall q$  ( $q = 1, \dots, m$ ) and  $R_{ij} = 0$  otherwise. Let also  $C_i = \sum_{j=1}^{n_S} R_{ij}$  and  $U_{ic} = 1$  ( $i = 1, \dots, n_R$ ;  $c = 1, \dots, C_i$ ) if  $t_i = t_c$  for categorical  $T$ , or  $t_c \in [t_i(1-p), t_i(1+p)]$  for continuous  $T$  and a precision of  $p \times 100\%$ , and  $U_{ic} = 0$  otherwise. We therefore denote by  $Z_i = \sum_{c=1}^{C_i} U_{ic}$  the total number of records that are a real match for unit  $i$  in  $D$ . Let  $I_i = 1$  if  $Z_i > 0$  and  $I_i = 0$  otherwise, and  $K_i = 1$  if  $C_i = 1 \wedge I_i = 1$  and  $K_i = 0$  otherwise. Disclosure risk can therefore be quantified by the following global measures:

- the expected match risk, given by  $\text{EMR} = \sum_{i=1}^{n_R} \frac{Z_i}{C_i}$
- and the true match risk, given by  $\text{TMR} = \sum_{i=1}^{n_R} K_i$

EMR reflects the chance of an user randomly establishing a true match for each unit  $i$  in  $D$  and TMR that of an user correctly and uniquely identifying each unit  $i$  in  $D$ .

### 3 Producing PUF for the Household Budget Survey

The Household Budget Survey (HBS) aims at producing data on consumption expenditure; its microdata is composed by records regarding households and household members. We generated as many household records as the number of households in the original sample. We began by simulating region (NUTS II) and household size by sampling from the corresponding estimated multinomial distribution, based on the relative frequency distributions in the original sample - we first simulated region and then household size, given region. For each synthetic household, a real household from the same region and size was randomly selected; the number of members in the synthetic household, as well as their sex and age, was taken to be that from the selected real household. Both approaches presented in 2.1 - Parametric and CART - were then used to generate the main identifying variables (country of birth, country of citizenship, marital status, level of studies completed, status in employment and economic sector in employment), as well as income and expenditure totals. In order to compare both approaches regarding the resulting disclosure risk, we generated 100 synthetic datasets from each approach, considering a random sample of 500 households from the HBS SUF to be our real data.

### 4 Results and discussion

Figures 1 and 2 illustrate respectively the distributions of two identifying variables and the total expenditure, obtained by generating a single synthetic data following each approach, in comparison with the real data (SUF). Good results were obtained regarding the main statistics computed from HBS data (e.g. the mean consumption expenditures of households (euros) - SUF: 20 391, Parametric: 19 942 and CART: 19 661 - and the at-risk-of-poverty rate (after social transfers) (%) - SUF: 14.8, Parametric: 19.2 and CART: 15.5). However, the additional flexibility from CART results in a slight increase in disclosure risk, as illustrated by figure 3.

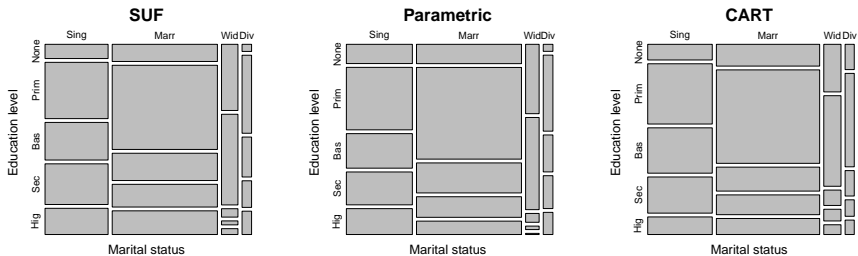


FIGURE 1. Weighted frequency distribution of education level by marital status, in the real (SUF) and synthetic HBS datasets.

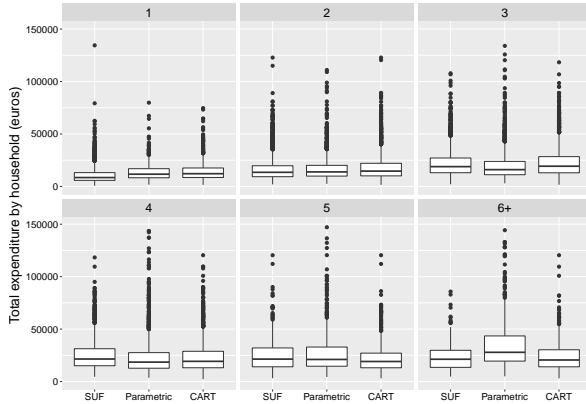


FIGURE 2. Total annual consumption expenditure by household size, in the real (SUF) and synthetic HBS datasets.

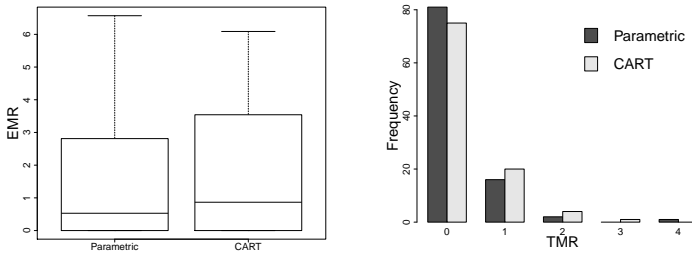


FIGURE 3. EMR and TMR distributions for 100 replications. Following the notation in 2.2,  $m = 6$  (sex, age, HH size, marital status, status in employment and country of citizenship),  $T$  are income and expenditure totals and  $p = 0.05$ .

## References

- Drechsler, J. and Reiter, J.P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, **55**, 3232–3243.
- Loong, B., Zaslavsky, A.M., He, Y. and Harrington, D.P. (2013). Disclosure control using partially synthetic data for large-scale health surveys, with applications to CanCORS. *Statistics in Medicine*, **32**, 4139–4161.
- Nowok, B., Raab, G.M. and Dibben, C. (2017). Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R. *Statistical Journal of the IAOS*, **33**, 785–796.
- Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, **27**, 85–95.



# Small Area Estimation for Land Use and Land Cover

Pedro Campos<sup>1</sup>, Suelma Pina<sup>2</sup>, A. Manuela Gonçalves<sup>2</sup>

<sup>1</sup> Statistics Portugal, Portugal

<sup>2</sup> University of Minho, Portugal

E-mail for correspondence: `pedro.campos@ine.pt`

**Abstract:** Small Area Estimation (SAE) is a part of statistical science that combines survey sampling and inference of finite populations with statistical modelling. The main objective of this paper is to analyze and test the implementation of different types of estimators of small domains in order to improve the quality of the estimates produced within the framework of the Farm Structure Survey (FSS) at NUTS III level. Under the EUROSTAT Land Use and Cover Area Statistical Survey (LUCAS) project, this is a fundamental tool for environmental studies, forestry and agricultural resource planning.

**Keywords:** Small Area Estimation; Regression Estimator; EBLUP; SEBLUP; Farm Structure Survey

## 1 Introduction

Nowadays, public and private institutions are increasingly seeking more detailed information to aid their decision-making process, and the National Statistical Offices do fall into this new paradigm. The need to produce reliable estimates for the total of variables of interest in small domains is fundamental. However, estimates cannot always be obtained through direct estimators (that use only the observations of the variable of interest belonging to the domain for the time period under analysis), because often there are no samples for these domains, or they are too small to obtain sufficient quality estimates. In order to solve this problem, several types of estimators for small domains have been proposed: some of them combine auxiliary information of the variable of interest in the domain and in different periods of time, or even consider variable sources of other domains (the so-called indirect estimators). The main objective of this paper is to develop, analyze and test the implementation of different types of small area

---

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

estimators in order to improve the quality of the estimates produced within the framework of the Farm Structure Survey (FSS) at regional (NUTS III) level. Currently, Statistics Portugal publishes these estimates at National (NUTS I) and Regional (NUTS II) levels. Under the EUROSTAT Land Use and Cover Area Statistical Survey (LUCAS) project, Statistics Portugal intends to use this information to detail the agriculture class, thus providing information on agricultural land use up to the third level of patent nomenclature in the Land Use and Land Cover Mapping (LULC), a fundamental tool for environmental studies, forestry and agricultural resource planning (EUROSTAT,2013). In this work, five different estimators (direct, modified and combined) are used to estimate 44 variables by NUTS III in mainland Portugal: the direct estimator (1 and 2), the estimator modified by the Regression, the EBLUP estimator using the Fay-Herriot method and the EBLUP estimator by the spatial level of the area (SEBLUP). Based on the results, we may conclude that when auxiliary variables are available, the estimator modified by the Regression performs better when compared to other estimators.

## 2 Small Area Estimators

In this section we introduce Small Area Estimation (SAE) and shortly describe the main estimators used in this work. In a stratified random sampling design, let  $U$  be a finite population of  $N$  distinct elements,  $U = \{1, \dots, N\}$ , the subpopulations (in this case, strata),  $U_h$ , with  $U_h \subset U, h = \{1, \dots, H\}$ , for which certain parameters have to be estimated according to the domain  $d$ . (see Figure 1).

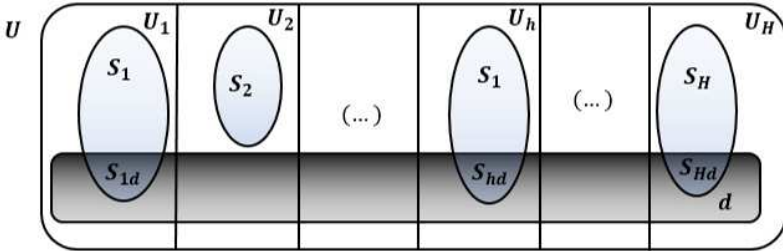


FIGURE 1. Representation of domains, under the SAE perspective

The population dimension of each stratum  $U_h$  is denoted by  $N_h$  with  $h = \{1, \dots, H\}$ , where  $N = \sum_{h=1}^H N_h$ , and the subpopulation dimension in  $U_{hd}$  is denoted by  $N_{hd}$ , where  $N_d = \sum_{h=1}^H N_{hd}$ ; we consider  $s$  as a sample

of size  $n$  collected from  $U$  that may be decomposed in  $s = \sum_{h=1}^H s_h$  and  $s_d = \sum_{h=1}^H s_{hd}$ , which are sampling units of size  $n_d$  and  $n_{hd}$  randomly selected, where  $n = \sum_{h=1}^H n_h$  and  $n_d = \sum_{h=1}^H n_{hd}$ .

We usually denote population  $U$  as being composed by two quantities,  $Y$  (the explained variable, or variable of interest) and  $X = (X_1, \dots, X_j) \in \mathbb{R}^j$ , the values of the covariates or auxiliary variables. Auxiliary variables are always assumed to be known, whereas the variable of interest may be unknown for some areas if individuals in these areas are not sampled. Assuming that we want to obtain estimates of the total,  $\tau_d$  the total of the variable of interest for the population of the domain of interest  $d$  is given by:  $\tau_d = \sum_{i \in U_d} Y_i$ .

In general, SAE models can be categorized in direct and indirect estimators. Direct estimators only consider the observations of the variable of interest belonging to the study domain for the time period under analysis, whereas indirect estimators take observations of the variable of interest as well as auxiliary sources outside the study domain for the considered period of time. The Model-based approach belongs to the class of indirect estimators and regression models are used here between data from the sample and auxiliary variables from other data sources, such as census and administrative records to "lend" information from similar areas (Rao and Molina, 2015). Indirect estimators can also be divided in synthetic and combined estimators which can be derived under a design-based approach or taking into account the fact that an explicit area level or unit level model exists. Combined estimators are basically weighted averages of a direct estimator and an indirect estimator (Rao and Molina, 2015, Pfeffermann, 2013).

## 2.1 Direct Estimators ( $D_1$ and $D_2$ )

We start with the fundamental Horvitz-Thompson estimator, defined in Rao and Molina (2015):

$$D_1 = \hat{\tau}_{D_1} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{i \in s_{hd}} y_i$$

$$Var(\hat{\tau}_{D_1}) = \sum_{h=1}^H \frac{N_{hd}(N_h - n_h)}{n_h} (s_{hd}^2 + (1 - \frac{N_{hd}}{N_h}) \bar{y}_{hd}^2)$$

A second estimator is used, where we assume to know the dimension of each population defined by the intersection of NUTS III with the strata defined a priori in the sampling plan: ( $N_{hd}$  e  $n_{hd}$ ):

$$D_2 = \hat{\tau}_{D_2} = \sum_{h=1}^H \frac{N_{hd}}{n_{hd}} \sum_{i \in s_{hd}} y_i$$

$$Var(\hat{\tau}_{D_2}) = \sum_{h=1}^H \frac{N_{hd}(N_h - n_h)}{n_h} s_{hd}^2$$

Where,  $s_{hd}^2$  is the sampling variance in the subsample defined by the intersection of stratum  $h$  with domain  $d$ .

## 2.2 Direct Estimator modified by Regression (Reg)

For the application of this estimator, it is necessary to know the values of the auxiliary variables for all units of the population at individual level, the vector of the totals of the auxiliary variables in domain  $\tau_{xd}$  and their observed values in the sample units of the subpopulation  $g, x_i, i \in s_g$ . The regression estimator for the total estimate is given by:

$$\hat{\tau}_{d,reg} = \hat{\tau}_d + (\tau_{xd} - \hat{\tau}_{xd})' \hat{\beta}_g$$

where  $\hat{\beta}_g$  is the estimator of regression parameters  $\beta_g = (\beta_{g1}, \beta_{g2}, \dots, \beta_{gp})'$ . In this case there is an implicit link model:  $y_i = x'_i \beta_g + \epsilon_i$ , with  $i \in U_g$

## 2.3 EBLUP and SEBLUP

The EBLUP is a combined estimator. Considering a finite population divided into  $D$  small domains, the Fay-Herriot base model (Rao and Molina, 2015) linearly relates the value of the  $d$ -th domain of the variable of interest  $\theta_d$  to a vector of  $p$  auxiliary variables aggregated at the  $x_d$  area level and includes an associated random  $v_d$  effect. The model is given by  $\theta = x'_d \beta + v_d, d = 1, \dots, D$ ; where  $\beta$  is a vector of regression parameters;  $v_d$  are the random effects. Then, the combined estimator SEBLUP,  $\hat{\theta}_{SEBLUP}$  of parameter  $\theta_d$  may be written as:

$$\hat{\theta}_{SEBLUP} = x'_d \beta + v_d + e_d = x'_d \beta + (I_D - \rho W)^{-1} u + e_d$$

The SEBLUP estimator considers a spatial component. The main difference between the two models (EBLUP and SEBLUP) lies in the fact that SEBLUP uses the information of the distances between the domains through a proximity matrix (Pfeffermann, 2013).

# 3 Data, Software, and Results

## 3.1 Data and Software

The Farm Structure Survey (FSS), also known as the Survey on the structure of agricultural holdings, is carried out by all European Union (EU) Member States and provides comparable statistics across countries and time, at regional levels (down to NUTS 3 level). The edition of 2013 considers more than 650 variables. In this study several strata has been considered, based on size class, area status, legal status of the holding, objective zone and farm type (INE, 2013). Therefore, the population has been divided in 765 strata, ( $h=1, \dots, 765$ ) and 23 domains or small areas, corresponding

to NUTS III, ( $d=1,...,23$ ). The overall population size ( $N$ ) is 236696 agricultural holdings and the sample size ( $n$ ) is 23108, representing about 9,76 % of the population. Algorithms to calculate the estimates, with the exception of the EBLUP estimator, were all programmed in R by the authors. The SEBLUP algorithm was obtained through the `eblupSFH` function of the R package `sae` (Molina and Marhuenda, 2013). In order to measure and compare the quality of the estimators, the coefficients of variation (CV) are computed and shown in percentage. To see if the spatial information introduced by the SEBLUP provided some improvement in the CV estimates, in the analysis of the results we also consider the results of the EBLUP estimator computed through the Fay-Herriot method ( $EBLUP_{FH}$ ).

### 3.2 Results

Results of the coefficient of variation (CV) of the five estimators are presented in Table 1.

TABLE 1. Results of the coefficient of variation (CV) of the five estimators

Estimator	CV range (%)	1st Quartile	Median	Mean	3rd Quartile	Quartile
$\hat{\tau}_{D_1}(Direct_1 or D_1)$	1.63-41.21	2.99	3.99	7.14	5.83	9.32
$\hat{\tau}_{D_2}(Direct_2 or D_2)$	1.29-18-82	2.12	2.57	3.72	3.84	3.61
$\hat{\tau}_{d,reg}(Reg)$	0.93-24.00	2.23	3.64	4.87	4.88	4.93
$\hat{\theta}_{SEBLUP}$	1.64-44.09	3.04	3.99	7.33	5.89	9.86
$\hat{\theta}_{EBLUP_{FH}}$	1.63-39.37	2.86	3.93	6.83	5.84	8.66

The wide variation of the CV range is due to the fact that different small areas (the NUTS III regions) differ much in terms of sample sizes. We can see (see Figure 2) that lowest values of CV were provided by Reg (the Direct Estimator modified by Regression) , although Direct 2 (the Direct Estimator 2) also performed well.

## 4 Conclusions

With regard to modified and indirect estimators Reg, SEBLUP and EBLUP, we found out that they present greater gains in precision when the sample size is larger and when the correlation between the dependent and independent variables is greater. When analyzing the CV estimates of the different estimators studied by NUTS III for one of the most important variables, UAA (Utilized Agricultural Area), the regions of Baixo Alentejo (184) and Alentejo Central (187) are the ones with the highest CV values when compared with those of the other NUTS III regions. This result ends

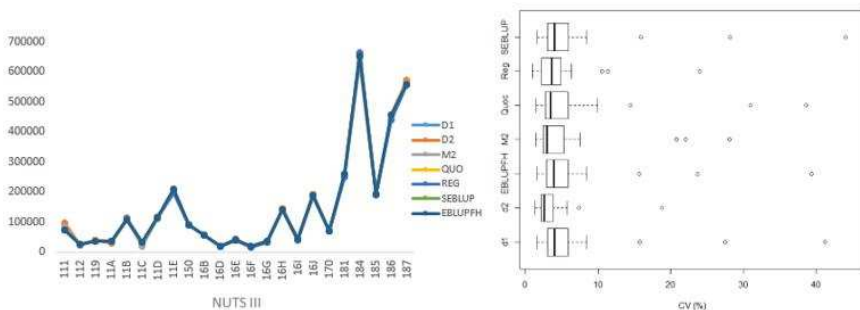


FIGURE 2. Graphical comparison of the estimates and boxplots of CV for the five estimators under analysis. (Note: we introduced two extra estimators: M2, the modified estimator and Quo, the Quotient estimator).

up harming the interpretation of the mean CV values of the estimators, since in general the CV estimates for the other regions are much lower.

## References

- EUROSTAT (2005). LUCAS 2009 (Land Use / Cover Area Frame Survey), *Quality report*. Luxembourg: Eurostat.
- Instituto Nacional de Estatística (INE) (2013). Inquerito a Estrutura das Explorações Agrícolas *Documento Metodológico*. Lisboa: INE.
- Molina, I., Marhuenda, Y. (2013). sae: An R package for Small Area Estimation In: *The R Journal*, 7, 1.
- Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, **28**, 1, 40, 40–68.
- Rao, J.N.K., Molina, I. (2015). Small Area Estimation, 2nd Edition *Wiley Series in Survey Methodology*, John Wiley and Sons, Inc., Hoboken, New Jersey