ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 21, Number 1, June 2017
Available online at http://acutm.math.ut.ee

Using k-anonymization for registry data: pitfalls and alternatives

Sten Anspal, Mart Kaska, and Indrek Seppo

Abstract. We describe an applied study of ICT students' employment in Estonia based on data from two national registries. The study offered an opportunity to compare results from both k-anonymised data as well as those from the novel Sharemind platform for privacy-preserving statistical computing, which offers a way to use confidential data for research without loss of information.

Comparison of results using k-anonymized and lossless data indicate substantial differences in estimates of students' employment rates. The results illustrate, on the basis of a real-world study, how the effects of k-anonymization can lead to considerable bias in estimates. While privacy-preserving computing does entail inconveniences because original microdata is not revealed to the statistician, this can be offset by greater confidence in the results.

1. Introduction

Like many countries, Estonia has collected large amounts of registry data in the course of provision of various public services. Covering many aspects of the society, it is potentially very useful for economic and social science research. In many cases, registry data are more up-to-date and cover more subjects than would be available from small surveys traditionally used for the purpose of such research, or even large-scale surveys carried out by national statistical offices. Naturally, registry data has its own limitations compared to surveys, since they have been collected for different purposes. Nevertheless, in many cases, registry data usefully complements survey data, or in some cases, is the only way to address a research question.

However, the challenge of using registry data is that their use is regulated by varying degrees of limitations related to confidentiality. This may be

Received September 30, 2016.
2010 *Mathematics Subject Classification.* 62P99.
Key words and phrases. Privacy-preserving computing; k-anonymization.
http://dx.doi.org/10.12697/ACUTM.2017.21.05

different from country to country: e.g., in some countries, income data is confidential, in others it is considered public. Even within a single country, different confidentiality rules apply to different datasets: e.g., in Estonia, companies' corporate registry data (balance sheet and annual report data) is public, while data on taxes on salaries paid by companies to individual employees is confidential.

Confidentiality requirements do not always mean there is no possibility of using registry data for research. After all, they typically apply to single data subjects (or groups in which single subjects can be identifiable), not to aggregate statistics that are the result of most research. There are different options researchers can use to address the confidentiality requirements, depending on the situation.

For example, in some cases signing a confidentiality agreement is considered a sufficient measure of data protection, in which case the researcher can use the data with relative ease for the purposes of research.

Sometimes, an additional condition is stipulated that the researcher is allowed to use the data only in a "secure room", a designated workplace on the premises of a trusted institution (such as the institution owning the data, or the national statistical office), which is not connected to the outside computer networks. However, sometimes this option is not available, for example because the legislation of the country does not allow such an arrangement, or because the data owner does not consider this a sufficient measure of protection.

Another possibility is to submit a query to the data owner, describing all the data analysis algorithms such as aggregations, summaries and statistical models, that are to be applied to the original registry data. This approach, however, places a great burden on the data owner – such a task is hardly ever a simple matter of running an algorithm on the original data but typically involves extensive work on cleaning and transforming the data. For this reason, this option is also not always available.

There is also the possibility to obtain explicit consent from data subjects for the use of data on them for research purposes. However, this involves significant costs in terms of money and time, given that the attraction of registry data is the possibility to use data that covers populations or large samples thereof. Also, this introduces the possibility that consent is withheld for some part of the population, resulting in a (possibly non-random) loss of observations.

Yet another possibility is to use k-anonymization. K-anonymity is a property of the data such that information on each subject in the dataset cannot be distinguished from at least k-1 other subjects. To illustrate this, a data set that included sex, age group and employment could be said to have 3-anonymity if there would be at least 3 observations for each combination of sex, age group and employment status. In this approach, observations for combinations of covariates for which there are less than 3 observations in the dataset would be dropped. This is the limitation of the k-anonymization approach: there is a trade-off between observation loss, number of covariates, and level of detail (number of categories in each covariate) in the data. This can result in a non-random loss of observations, with implications on statistical inference.

In this paper, we descibe a real-world applied study based on combined data from two Estonian public registries, using two different approaches. One of the approaches used was 3-anonymization, the other was a novel technological solution, the SHAREMIND platform for privacy-preserving statistical computation (see [2], [7]). These two approaches and the results and limitations of the two approaches are described. The main aim of this paper is to highlight the limitations of using k-anonymization for research on registry data, and introduce the SHAREMIND platform as a viable alternative.

In the following section, we describe the study and the two approaches used. Section 3 describes the data, followed by a discussion of the results in Section 4, Section 5 concludes the results.

2. Methods

2.1. The research question of the applied study. The objective of the applied study was to examine the relationship between IT students' graduation and employment during studies, using data from national registries. The motivation for the study is the low graduation rate among Estonian ICT students. The problem has been commonly attributed to the fact that ICT companies have offered lucrative jobs to students, making timely graduation more difficult. It is certainly the case that ICT skills are in high demand in the labour market and that ICT students work during their studies (employment during studies is common among students in Estonia in general), the prevalence of this phenomenon is not known. The research question was proposed by and the study was carried out for the Estonian Association of Information Technology and Communications, an Association of ICT companies and other organisations with focus on ICT.

The speficic research questions addressed in the study were the following:

- What is the share of students who graduate in time among ICT students;
- What is the employment rate among ICT students;
- What is the share of students working in ICT companies among working ICT students.

For the sake of brevity and because of the focus of this paper, we present here only the results for the second research question, for students enrolled in 2006 on the bachelor level, since this is sufficient to illustrate the two approaches to privacy protection used.

In the study, we examined ICT students' employment and graduation using data from two national registries: the Estonian Education Information System and the Tax and Customs Board's Register of Taxable Persons (the data used is described in Section 3) and two methods for preserving privacy: k-anonymization and the SHAREMIND platform for privacy preserving statistical computation. The comparison of working processes and results obtained using these two approaches was the second objective of the study. The two approaches will be discussed in the following sub-sections.

2.2. *K*-anonymization. Since data on declared taxes from the Estonian Tax and Customs Board used in the study was confidential, it was necessary to ensure that no single person's tax information could be directly or indirectly revealed to the researchers. For this purpose, 3-anonymisation was used: persons with characteristics such that the speficic combination of characteristics was found in less than two other persons were removed from the sample. Thus, calculations were not based on the population of persons in the database but on a sample, since persons with rare combinations of personal characteristics were left out. This method of anonymization was chosen because it was a previously established practice accepted by the Tax and Customs Board.

K-anonymization has been previously used in analysis of Estonian registry data. In [1], an evaluation of labour market training was carried out, using Estonian Unemployment Insurance Fund and Tax and Customs Board data and 3-anonymization. Using coarsened exact matching and survival analysis, the study found that labour market training raised the probability of employment by 6% during the first year after completion of the training. Approximately at the same time, the same research question was addressed in a different study (see [6]) that had privileged access to the same data, without the need of k-anonymization and thus no loss of observations. Remarkably, the results obtained in that study were qualitatively as well as quantitatively similar. However, as will be demonstrated below, this cannot always be expected to be the case.

Note that the k-anonymization technique used here is very restrictive since all cells with less than k observations are simply dropped. An alternative would have been to merge cells until at least k observations are achieved. For example, cells for curricula "x" and "y" with 2 observations each could be merged into a cell "z", containing 4 observations and thus retaining data on other covariates than the respective curricula (which would be labeled as missing values). In such a case, the loss of data would have been smaller and the implications less drastic than described below. The more restrictive approach was used because a number of different research questions were posed in the study. It was preferred to use a single dataset with no missing values for all research questions, so that all estimates would be reported on the basis of identical data. However, this goal entailed a substantial tradeoff in terms of data loss and in hindsight, bias in estimates.

It is easy to see that creating groups of at least three persons based on their characteristics leaves out the more people, the more characteristics are used to define such groups. For example, if one considers people who enrolled in the Computer Systems curriculum at Tallinn Technical University in 2006 and graduated in 2009, it is quite likely that at least three people could be found that have these characteristics. If, however, one would like to consider also sex and age, it becomes less likely that there are at least three females aged 20−24 among the group of people who studied computer systems at TTU in 2006−2009. Thus, the greater the number of characteristics, and the greater the number of categories in each characteristic (such as smaller age ranges), the greater the loss of observations.

The first challenge is to find a compromise between the number and detail of characteristics used in the study and the number of observations left in the sample. In this study, the following students characteristics were used:

- Study level (BA, MA, Ph D.)
- Curriculum group code and name
- Curriculum code and name
- Name and code of school
- Year of admission
- Whether the student graduated in time
- Year of graduation or dropping out
- Nominal length of study years in curriculum

Initially, also age, sex and form of study (full-time, non-stationary study) were considered, but were omitted because of excessive loss of observations.

It should be noted that the choice of covariates and their level of detail was greatly simplified by the fact that data from the Educational Information System were not private from the point of view of the researcher, only the data from the Tax and Customs Board were. Had both datasets been confidential, the optimal choice of covariates would have been a nearly impossible task.

The number of observations in the original population and after anonymization are reported in Table 1. The loss of observations is greater for non-ICT students because of the greater number of curricula with few students, for which the probability of small cells occurring was higher.

This method of anonymization entails a number of limitations. Foremost of these is that persons whose combination of characteristics is found in less than 3 people are omitted. Moreover, the sample that remains after that omission is not random, thus estimates may be biased. Also, characteristics that had to be omitted cannot be included in the study. The time of admission and graduation or dropping out is imprecise (year). This introduces a potential

Table 1. Observations in the population and in the sample obtained after 3-anonymization, BA students by study subject.

Study subject	Population	Sample	Loss, %
ICT	4,248	3,878	9%
Non-ICT	53,113	45,527	14%

error in estimating employment rates: if only the calendar year is used, it is not certain whether, e.g., employment in the first year of study should count as employment during studies or not (although most BA admissions take place in the fall semester, a few are admitted during the spring semester).

These limitations were not present when using the Sharemind technology described in the following subsection, with which there is no necessity to omit persons with rare combinations of characteristics.

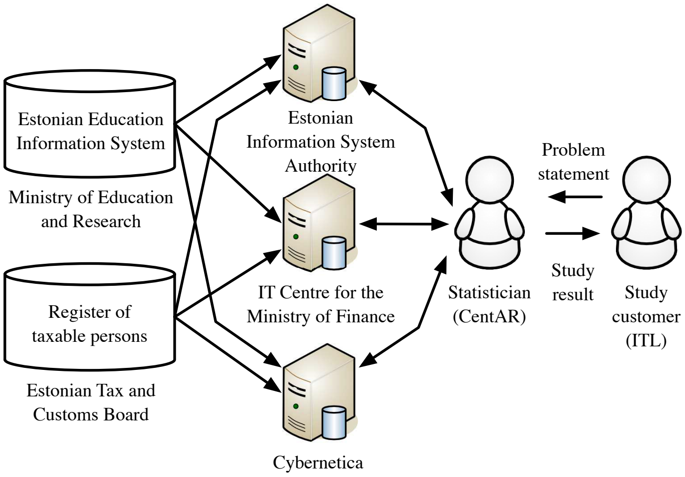
2.3. Sharemind. In this section, we give a brief non-technical overview of the Sharemind platform that was used in this study. For technical details, see [2].

Sharemind is a secure computation framework that implements secure multiparty computation (MPC), a cryptographic method for securely processing data among several parties. With secure multiparty computation, functions over input data are computed jointly by several different parties, such that input data remains private to anyone other than the input party, i.e., the owner of the data. The input party uses secret sharing on their private input, i.e., an algorithm is used to split the original private data into a number of random pieces, called shares (in the implementation of Sharemind used in this study, the number of shares was three). Any observation in the original data can only be reconstructed from all three shares, not from any one or two shares.

There are three computing parties who run instances of the Sharemind application to perform computations on the shares on their servers. The statistician (the result party) uses a client application to run statistical algorithms that are executed on the shares by the computing parties, without the private inputs being revealed to either the statistician or the computing parties. Only the non-private results of the computations are displayed to the statistician.

The Sharemind framework is programmable: applications can be developed in the SecreC programming language and executed securely on the platform. In addition to SecreC, there is the Rmind tool (see [4] and [5] for more details) designed to make it easier to carry out data transformation and statistical operations. Rmind has a syntax similar to the R programming language; the statistician enters commands on the command prompt (or runs

FIGURE 2.1. Stakeholders in the statistical study.


Estonian Education Information System
Ministry of Education and Research
Register of taxable persons
Estonian Tax and Customs Board
Estonian Information System Authority
IT Centre for the Ministry of Finance
Cybernetica
Statistician (CentAR)
Problem statement
Study result
Study customer (ITL)

Reproduced from [3].

scripts) in the client application, which are then executed as pre-compiled SecreC applications on the three servers of the computing parties. As of the time when the study was carried out, the RMIND tool supported a number of statistical computations, including various descriptive statistics, common statistical tests, the general linear model, graphing, and data transformation utilities (the algorithms are described in more detail in [4]). Private data is never displayed to the statistician, aggregate results are only displayed to the user if they have been calculated on at least three observations.

3. Data

To answer the research question described in the previous section, combined data from the Estonian Education Information System and the Register of Taxable Persons were used. Data on students' admission, study, graduation and dropping out were obtained from the Estonian Education Information System at the Estonian Ministry of Education and Research. Data on students' employment and pay were obtained from the Register of Taxable Persons at the Estonian Tax and Customs Board.

The query from the Estonian Education Information System included persons who were admitted in Estonian higher or applied higher education institutions in 2006 or later or who graduated or dropped out in 2005 or later. Since we wanted to compare ICT and non-ICT students, data for all curricula was used. The fields included in the original query are reported in Table 2.

TABLE 2. Fields in the original datasets.

Estonian Education Information System	Tax and Customs Board
ID	**ID**
Sex	Employer ID
Year of birth	**Year (2004 onwards)**
Level of study	**Month**
Curriculum group	**Income taxable with social tax in the month under consideration**
Curriculum	Employer's NACE code (industry)
School	Whether the employer is a member of the Estonian Association of Information Technology and Communication (logical variable based on list of employer codes submitted to the Tax and Customs Board)
Year of admission	Employer's annual average number of employees
Status as of Nov 10 of given year (2005 onwards)	Social tax paid on grounds other than labour income
Graduation within nominal study time (logical variable)	
Form of study at admission	
Year of graduation/dropping out	
Nominal study time for curriculum (years)	
Nominal study time for curriculum (months)	

Note: Not all fields in the original datasets were eventually used in calculations described in this paper. Fields in bold type indicate fields that were used.

This dataset was not private to the researcher. Persons' names and national IDs codes were not included (the ID used in the dataset was a pseudo-ID code generated for the purposes of this study), but otherwise all values of all observations in the dataset were available to the researcher.

Since the Estonian Education Information System does not include data on employment and pay, it was combined with data from the Register of Taxable

Persons, which was confidential and needed to be anonymized.[1] The fields included in the latter dataset are listed in Table 2.

The study plan was submitted for approval to the Estonian Data Protection Inspectorate. Since the study involved the use of a novel technology, which had not been previously used on data from Estonian national registries, it necessitated a lengthy review of the technology and processes involved. The Inspectorate's assessment concluded, "Based on your application and supplements we conclude that in the described study no processing of personal data or sensitive personal data will be taking place". Thus, it was admitted that the SHAREMIND technology offers sufficient privacy protection, and that the processing of shares by the three computing parties did not constitute processing of the original data.

The data owners used the SHAREMIND import tool to secret share the data and upload the shares to the three computing parties. The data were then transformed from their original format (long format) into the form necessary for carrying out statistical calculations (wide format), and merged. Since this process was carried out in SHAREMIND, it had to be done without the researchers having access to individual observations, which naturally poses special challenges. This extract, transform and load (ETL) process and the verification of the results was prepared using test data by researchers at Cybernetica and is documented in detail in [3]. The ETL was then performed by the authors of this paper.

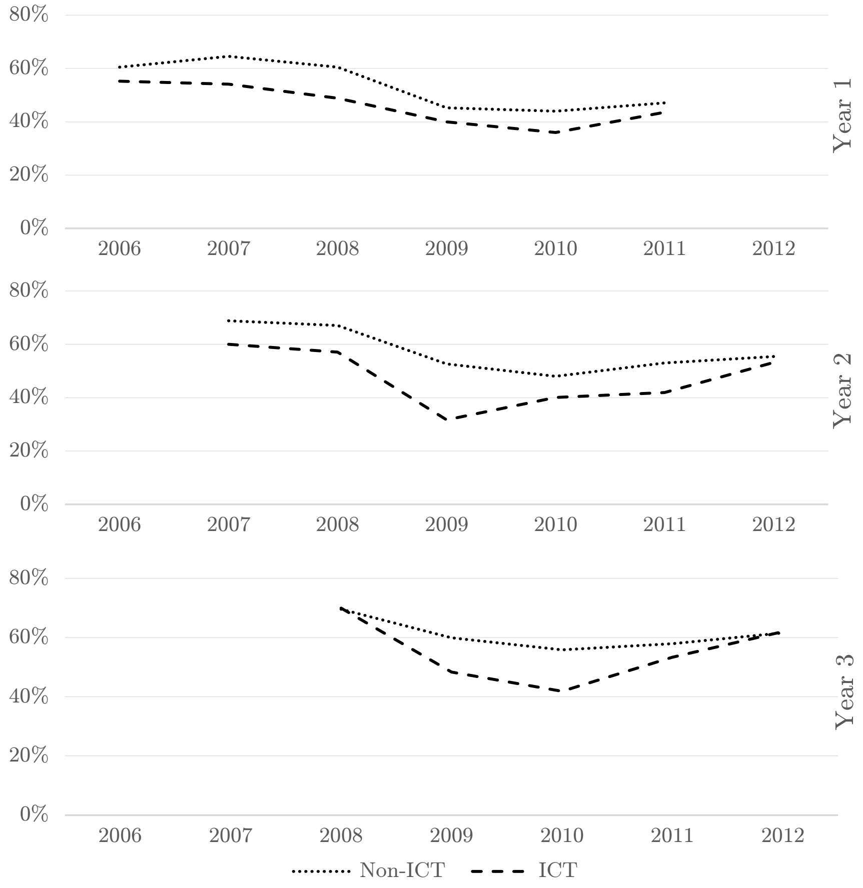
Before applying the algorithms for data transformation and calculation of results on the actual data on the SHAREMIND platform, the calculations were performed on a non-confidential dummy dataset in both the SHAREMIND platform and in Stata software jointly by Cybernetica and the authors. The dummy dataset was constructed on the basis of metadata.

4. Results

Figures 4.1 and 4.2 present the results regarding employment rates of the first, second and third-year non-ICT and ICT students in bachelor level curricula from the k-anonymized and SHAREMIND studies, respectively. Qualitatively, the results are similar: employment rates for the latter group are in fact lower than for non-ICT students in most years under consideration. This runs counter to the hypothetical explanation for the lower graduation rates of ICT students that was tested in the study, namely that ICT students' higher employment is the culprit behind their lower graduation rates.

[1] The query from the Register of Taxable Persons was made on the basis of the list of national IDs transmitted to the Tax and Customs Board by the Ministry of Education and Research, so that only data for people included in the study, i.e., higher education students during the time period under consideration, were included in the query.

FIGURE 4.1. Results using 3-anonymized data: employment rates of non-ICT and ICT bachelor level first-, second- and third-year students by year, 2006-2012.



There are also years in which ICT students employment rate exceeds that of non-ICT students', but those are in the minority.

In some years, the differences in employment rates are remarkably different, in excess of 10 percentage points. In particular, the Great Recession of 2008−2010 reduced employment rates for all students, but especially so for ICT students: employment rates for third-year students dropped from about 70% to less than 50%.

However, in terms of the comparison between the calculations obtained by using 3-anonymised data and SHAREMIND, the qualitative similarity between the results presented in the Figures 4.1 and 4.2 is misleading. In terms of quantitative estimates of employment rates, the results are quite different.

Figure 4.2. Results using lossless data with Sharemind: employment rates of non-ICT and ICT bachelor level first-, second- and third-year students by year, 2006-2012.

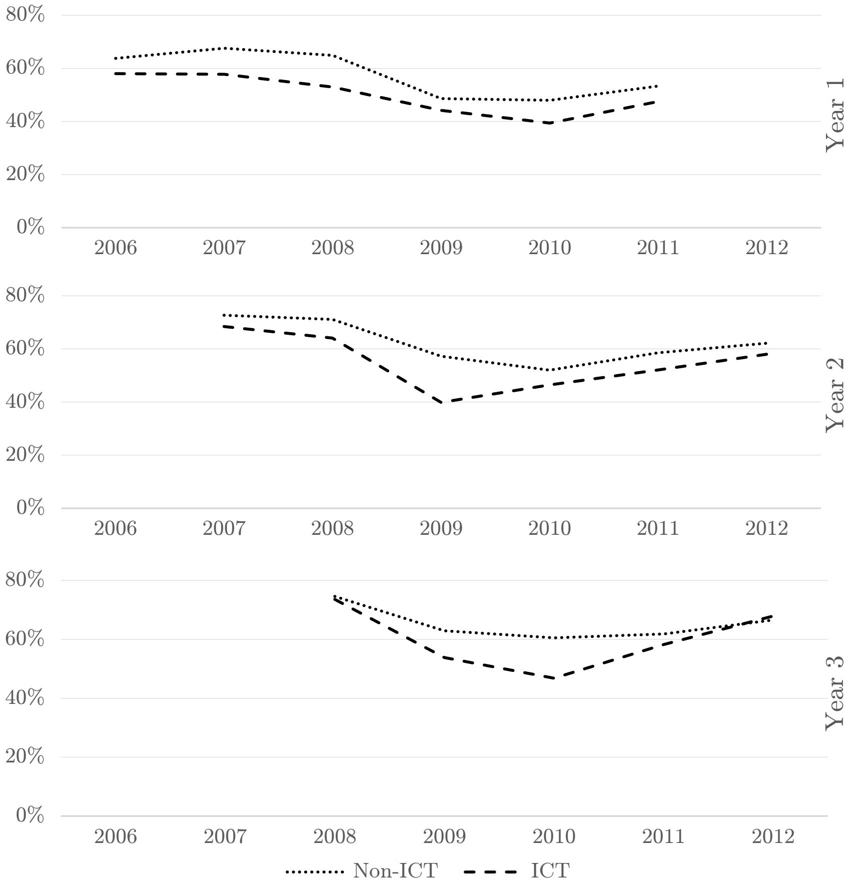
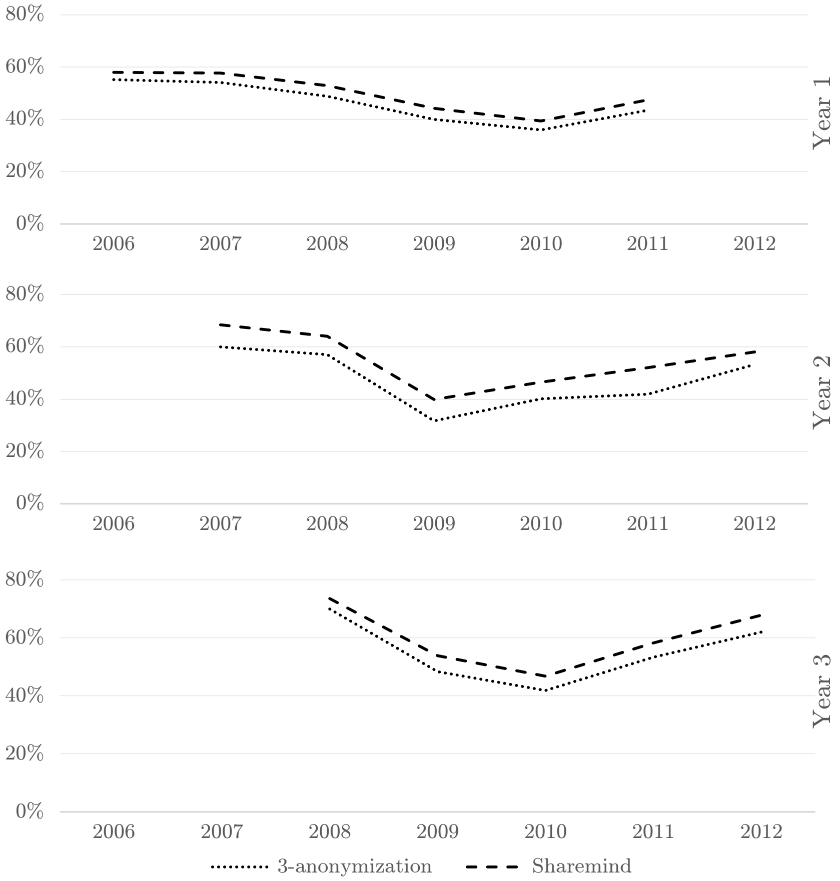


Figure 4.3 juxtaposes ICT students' employment rate estimates obtained using 3-anonymization and using Sharemind. The results indicate striking differences: for example, in 2009, the employment rate for second-year students was estimated at 32% using 3-anonymized data but at 40% using lossless data with Sharemind. This is a difference of 20%; an error of this magnitude would make the result difficult to use for policy decisions.

What compounds the problem with that estimation bias is that it is hard to quantify: since the sample drawn from the population in the process of 3-anonymization is non-random, basic methods for estimating confidence intervals are not applicable. Furthermore, it is not the case that we are more likely to observe certain combinations of characteristics than others and are

Figure 4.3. Comparison of results from 3-anonymized and lossless data.



able to correct for it by applying weights. Rather, some combinations of characteristics are not observed at all in the sample, even though they are present in the population. This makes it difficult to correct for the bias. It is possible that a different setup for k-anonymization (use of different covariates, different level of detail) would have yielded better performance in terms of inference. However, since querying national registries incurs costs to the data owners in terms of time and money, the researcher has typically only a single chance to formulate the query, so the setup has to be formulated in advance.

5. Conclusion

In this paper, we have described the results from an applied study of the ICT students' employment in Estonia based on data from two national registers. The study offered an opportunity to compare results from both k-anonymised data as well as those from privacy-preserving computations on the SHAREMIND platform, which offers a way to use confidential data for research without loss of observations.

The results provide a real-world example of how the effects of k-anonymization can be drastic and unpredictable in terms of inference. The comparison of results indicated large differences in estimates of employment rates obtained using 3-anonymization compared to those obtained using SHAREMIND, without loss of observations. The implication is that using 3-anonymized confidential data for research should only be done with great care, e.g., in situations in which it is possible to keep the loss of observations to a minimum. Since SHAREMIND involves no loss of information, it is the superior option in terms of statistical description of the population in cases where the compromises involved in k-anonymization are unacceptable.

However, SHAREMIND involves its own limitations, the most obvious one being that individual values in the original data are not revealed to the statistician. This is certainly an inconvenience in terms of cleaning the data, identifying data input errors or irregularities, and validating the results of transformations or statistical algorithms. These limitations can be overcome, but this requires changes in the statistician's usual workflow: the scripts used in the data analysis must incorporate extensive procedures of validation in order to explicitly test for any errors in data input or computations. Using non-confidential dummy datasets (or k-anonymised real data) to validate data transformation operations can be very helpful. However, great care should be taken that these validation procedures would be sufficiently exhaustive to exclude any reasonable doubt in data input or transformation errors. This necessitates budgeting more time for carrying out the study compared to traditional approaches. Another limitation of SHAREMIND is its limited range of statistical methods that have been implemented to date. However, since SHAREMIND is undergoing rapid development, this may change in the future.

Although privacy-preserving computations entailed time costs related to unobservability of individual observations and therefore additional efforts to validate the computations, these can be offset by greater confidence in the results.

Acknowledgments

The authors thank, without implicating, the Estonian Data Protection Inspectorate, the Information System Authority, the Ministry of Education and

Research, the Information Technology Centre of the Ministry of Finance, the State Information System Agency and the Tax and Customs Board for their collaboration. The study would not have been possible without the crucial participation of partners in the project from Cybernetica: Dan Bogdanov, Liina Kamm, Baldur Kubo, Ville Sokk, and Riivo Talviste. The authors also thank the anonymous referee for helpful comments and suggestions.

References

[1] Sten Anspal, Janno Järve, Kristel Jääts, Epp Kallaste, Kirke Maar, Annemai Mägi, and Anna Toots, *Evaluation of the Services of the Programme "Increasing the supply of qualified labour 2007–2013" : interim evaluation of the services of wage subsidy and labour market training*, Research report for the Estonian Ministry of Social Affairs, Estonian Centre for Applied Research (CentAR) and InterAct Projektid & Koolitus, 2012, 199 pages. (Estonian)

[2] Dan Bogdanov, *Sharemind: Programmable Secure Computations with Practical Applications*, PhD thesis, University of Tartu, 2013.

[3] Dan Bogdanov, Liina Kamm, Baldur Kubo, Reimo Rebane, Ville Sokk, and Riivo Talviste, *Students and taxes: a privacy-preserving study using secure computation*, Proceedings on Privacy Enhancing Technologies No. 3 (2016), 117–135.

[4] Dan Bogdanov, Liina Kamm, Sven Laur, and Ville Sokk, *Rmind: a tool for cryptographically secure statistical analysis*, IEEE Transactions on Dependable and Secure Computing **PP** (2016), 14 pages. doi: 10.1109/TDSC.2016.2587623.

[5] Liina Kamm, *Privacy-preserving statistical analysis using secure multi-party computation*, PhD thesis, University of Tartu, 2015.

[6] Anne Lauringson, Kristi Villsaar, Liis Tammik, and Teele Luhavee. *Impact assessment of labour market training*, Estonian Unemployment Insurance Fund, 2011, 97 pages. (Estonian)

[7] Riivo Talviste, *Applying Secure Multi-party Computation in Practice*, PhD thesis, University of Tartu, 2016.

Nunne 5-3, 10133 Tallinn, Estonia
E-mail address: sten.anspal@centar.ee

Tedre 55-37, 13417 Tallinn, Estonia
E-mail address: mart.kaska@centar.ee

Kaupmehe 6-62, 10114 Tallinn, Estonia
E-mail address: indrek.seppo@centar.ee