www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 4, No: 3, 2017 (319-333)

El-Cezerî Journal of Science and Engineering
Vol: 4, No: 3, 2017 (319-333)

ECJSE

Research Paper / Makale

Statistical Methods of Confidentiality for Micro Data and Developing an Object Oriented Statistical Disclosure Control Software

Abdulsamet HAŞILOĞLU[1], Abdulkadir BALI[2]

[1]Department of Computer Engineering, Atatürk University, Erzurum, 25240, Turkey
[2]Ataturk University, Institute of Science and Technology, Computer Engineering, Erzurum, 25240, Turkey
e-mail: asamet@atauni.edu.tr[1], abdulkadir.bali@tuik.gov.tr[2]

Received/Geliş: 19.05.2017 **Revised/Düzeltme:** - **Accepted/Kabul:** 30.05.2017

Abstract: Statistical offices collect large amounts of data for statistical purposes. A basic principle in statistical legal frameworks is that data collected for statistical purposes may only be used for the production of statistics. However, statistical offices experience increasing pressure from scientists and governments to provide access to detailed data. There are high costs and risks associated with micro data access. When micro data sets are released, it is possible that external users may attempt to breach confidentiality. In this paper, an object-oriented statistical disclosure control software, OOSDCS was developed to facilitate statisticians and to apply statistical disclosure control methods to create safe micro data files. The developed hybrid, flexible and interactive software was successfully applied as a disclosure control method.

Keywords: Statistical Confidentiality, Statistical Disclosure Control, Disclosure Risk, Micro Data, Software.

Mikro Verilerin Gizliliğinin İstatistiksel Yöntemleri ve Bir Bireye Yönelik İstatistiksel Açıklamaların Kontrol Yazılımını Geliştirme

Özet: İstatistik ofisleri, istatistiksel amaçlar için büyük miktarda veri toplamaktadır. İstatistiki yasal çerçevelerdeki temel ilke istatistiksel amaçlar için toplanan verilerin yalnızca istatistik üretimi için kullanılabilmesidir. Bununla birlikte, istatistiksel bürolar, bilim insanlarının ve hükümetlerin detaylı verilere erişim sağlamak için gittikçe artan baskıyı yaşıyor. Mikro veri erişimi ile ilişkili yüksek maliyetler ve riskler vardır. Mikro veri setleri serbest bırakıldığında, harici kullanıcıların gizliliği ihlal etmeye çalışması olasıdır. Bu yazıda, istatistikçileri kolaylaştırmak ve güvenli mikro veri dosyaları oluşturmak için istatistiksel açıklama kontrol yöntemleri uygulamak için nesne yönelimli istatistiksel açıklama kontrol yazılımı OOSDCS geliştirilmiştir. Geliştirilen hibrid, esnek ve interaktif yazılım, bir açıklama kontrol yöntemi olarak başarıyla uygulanmıştır.

Anahtar kelimeler: İstatistiki Gizlilik, İstatistiki Bilgi Kontrol, Bilgilendirme Riski, Mikro Veri, Yazılım

1. Introduction

Having high quality data is necessary to advance science and improve living conditions. Reliable data sources on human behavior allow researchers to answer complicated questions and to guide political/antipolitical decisions. However, as these data presented for public use carry individual characteristics, electronic tools to unify and analyze this data create a large danger for individual

How to cite this article
Haşıloğlu, A., Balı, A., "*Statistical Methods of Confidentiality for Micro Data and Developing an Object Oriented Statistical Disclosure Control Software*" El-Cezerî Journal of Science and Engineering, 2017, 4(3); 319-333.

Bu makaleye atıf yapmak için
Haşıloğlu, A., Balı, A., "*Mikro Verilerin Gizliliğinin İstatistiksel Yöntemleri ve Bir Bireye Yönelik İstatistiksel Açıklamaların Kontrol Yazılımını Geliştirme*" El-Cezerî Fen ve Mühendislik Dergisi 2017, 4(3); 319-333.

data protection. Statistical confidentiality control researches methods to keep statistical data contents secret and prevent publication of information that may relate assets to certain individuals.

In developed countries trade confidentiality laws audit this type of sharing [1]. This approach generally masks information belonging to individuals that generally determines identity. However, protecting data confidentiality has become more difficult in recent times. New information forms such as video data, biological samples, information movements and human behavior do not have a tabulated form. However, this micro data may be publicized after information is anonymized. Anonymization of statistical research is generally insufficient and with the correct skills and a computer, anyone with access to public data has been shown to cause confidentiality breaches [2]. Statistical disclosure control is a combination of administrative, legal and information technology with the aim of defining an appropriate data publication strategy based on a risk management approach. Different products in a data publication strategy may require different applications of statistical disclosure control and creation of different tools. Currently better statistical tools are being developed to present data in more appropriate fashion [3]. These tools are primarily helpful to appropriately assess risks. Secondly, they transform confidential data into harmless form to be shared publicly and are helpful to present it as tabulated data or anonymized micro data.

A study by Hundepool et al. used audits such as output from analyses like Tabular Data Protection, Dynamic Data Bases, Micro Data Protection and Statistical Analysis Output Protection to prevent possible confidentiality disclosures. This is a developing area within statistical disclosure control research.

2. Statistical Disclosure Control

Here, while the data provider of the National Statistical Office manages confidentiality risks, the type of approach for application with the aim of fulfilling the needs of data users at the highest level is explained. Hundepool et al. defined a general process of five steps for the problem of protecting confidentiality in production of statistical data [4].

These steps are;

- Evaluation of whether there is a need to protect confidentiality,
- Determination of characteristic properties carried by the data,
- Deciding which disclosure risks the data needs to be protected against,
- Choice of disclosure control methods,

Implementation.

During the disclosure control process, in addition to transparency, for assessment decisions taken and all risk determination processes are documented. Users need to be aware of processes the data set has undergone for disclosure risk and the applied control methods. Linked to implementation of disclosure control methods, users should be provided with information on what degree of change has been made to the data set by any method. Additionally users should be informed about the techniques used with the condition that they will not recreate the undisclosed data.

2.1. Reasons for Protecting Privacy

There are three basic reasons relating to the necessity of protecting confidentiality in the production of statistical data. These are international agreements, national laws and directives and ethical principles. To ensure protection of confidentiality of statistical products, it is necessary to know the characteristic properties of the data very well, because these properties affect both the disclosure risk and the choice of appropriate disclosure control method. Examples of characteristic properties of data include information such as whether data is compiled as a full inventory or as a sample, the non-response rate and scope of the survey, are variables categorical or continuous, and whether output will be micro data, size tables or frequency tables. Statistical publications should primarily be prepared according to the needs of the users. In situations where there may be disclosure risk, it is possible to apply a method combining the above understanding. While it may not be possible to fully remove all risks carried by data, risks may be minimized to an acceptable level. These precautions should be applied with statistical disclosure control or by limiting use of outputs. It is necessary to ensure a balance between many factors in the approaches to precautions. Using information loss measurements to measure the effects on usefulness of information after precautions determines which approach is more useful. A balance should be created between information lost by the methods used and the possibility of disclosure of personal information. The final stage in the disclosure control process is implementing the control and publishing the statistics. This process requires development of computer software. The software should present an appropriate interface allowing choice of disclosure control method and setting of parameters belonging to this method. While the most important aim is to protect confidentiality, the software should use available resources, take an acceptable amount of time, and ensure open, consistent and practical solutions.

3. Materials and Methods

3.1. Statistical Disclosure Control Methods

If a micro data file is not considered confidential, it requires changes to be made to obtain a confidential file. These changes generally are completed using statistical disclosure control. Examples of statistical disclosure control methods include global recoding, local cloaking, post randomization (PRAM), adding noise, micro aggregation and top and bottom value coding. When applied to micro data these techniques reduce the amount of information transmitted. Expressing variables as lower values or in more general categories, hiding values in the records or exchanging with other values may be shown to be causes of information loss.

3.1.1. Global Recoding Method

Global recoding is coding variables under many categories into a single category. An example of global recoding is creation of occupational groups from occupations, or representing settlement units with the name of larger settlement units. Global recoding is not just applied to the non-confidential portion of the data set but to the whole of the data set. The reason for this is that it is necessary that each variable in the data set have an unchanging categorical structure within itself.

3.1.2. Local Cloaking Method

Records for a variable with local cloaking applied is changed with an unknown value from non-confidential values (generally marked *). Local cloaking can be applied to a single variable, but can also be applied to more than one variable. Choosing the lowest number of records requiring local cloaking makes it possible to reduce the amount of information loss.

3.1.3. Top and Bottom Coding Method

While global recoding is a technique that may be applied to all categorical variables, top and bottom coding can only be applied to ordered categorical variables (age group, educational level, income level, etc.). Top coding evaluates a certain categorical value and all values above it within a single category (e.g., 95+ years). Bottom coding is the inverse, with a certain categorical value and all values below it shown in a single category. For top and bottom coding while determining which values will be upper and lower limits, the aim is to create new categories by combining as few other categories as possible and containing as few records as possible to balance information loss and disclosure risk. In other words, the upper threshold value should be chosen as high as possible and the lower threshold value should be chosen as low as possible without violating confidentiality limits.

Top and bottom coding can be applied to continuous variables, in addition to categorical variables. It is sufficient that values taken from variables be linearly ordered (e.g., monthly net income less than 1300 TL). In practice as statistics are known to be obtained from assemblages containing finite elements, there is no difference between a continuous variable and a categorical variable with multiple categories. After continuous variables with real number values are divided into categories, they may be subjected to top and bottom coding (e.g., babies with birth weight below 1000 g).

3.1.4. Post Randomization Method: PRAM

PRAM is a disclosure control method applied to categorical variables [5],[6]. When PRAM is applied the value of at least one categorical variable in each record in a micro data file is changed. This is completed independently of other records linked to a previously determined probability mechanism. In the end, the original file is changed to make it more difficult for badly intentioned individuals to match records in the file with real individuals in the society. As the probability mechanism is known when PRAM is applied, it is still possible to use the mixed data to re-estimate original data. The theoretical explanation of PRAM is as follows; Let the variable which PRAM will be applied to in the original micro data file be shown as ξ, and the equivalent of the same variable in the mixed file be shown as X. Additionally let ξ and X have K categories between 1..K. In PRAM the possibility of k value switching with l value is shown by the transition probability $p_{kl}=P(X=l|\ \xi=k)$. PRAM represents a KXK matrix. When applying PRAM, the record r with value k in the original file is switched with a value taken from the probability distribution p_{k1},.., p_{kk}. This procedure is completed for each record in the original file independent of other records [5],[6], with p_{kk} the probability that the record with value k in the original file remains unchanged. As PRAM is a method based on randomness when the same PRAM matrix is applied to the same original file at a different time, different results will be obtained. If the transition probability is known, unbiased comparison tables are easily obtained. For complicated analyses, to improve data mixing by PRAM detailed methods are required.

Another method similar to PRAM is the randomized response method [7]. The randomized response method applies probability methods to very sensitive questions with low probability of a correct response by respondents in face-to-face interviews with the aim of preventing the user being identified by the pollster. The randomized response method is applied before the survey, while PRAM is applied after the survey.

3.1.5. Microaggregation Method

Microaggregation is a statistical disclosure control method applied mainly to quantitative micro data. Microaggregation uses the philosophy that records comprising k or more units with no unit dominant belong to clusters, in accordance with the confidentiality rules for publication of statistical data. Before publication, instead of values of units, the mean calculated for micro groups of units comprising k or more numbers is inserted [8].

4. Software

4.1. Object Oriented Statistical Disclosure Control Software (OOSDCS)

A variety of software has been developed to apply confidentiality and statistical disclosure controls on micro data. The most important of these softwares are Mu-Argus, R/sdcMicro and SUDA. Mu-Argue was developed under the auspices of the European Union-sponsored CASC (Computational Aspects of Statistical Confidentiality) project [4]. It offers the statistical disclosure control methods of local cloaking, global recoding, cutting, top and bottom coding, microaggregation, adding noise, PRAM and rank exchange. It has a graphical user interface. *R* is an open-source code mathematical software while sdcMicro is the statistical disclosure control application packet developed for micro data in *R* [9]. R is in C/C++ languages with sdcMicro and other packets written in a type of interpretative language for R. The program presents command prompts when commands in the assumed R language are entered. It is possible to create program threads with graphical user interface for R language. SUDA (Special Unique Detection Algorithm) is a program allowing determination of axiomatic risk carried by micro data at the record level. Its use is limited to population census data [10].

The *OOSDCS* software program developed within this study presents the methods of local cloaking, global recoding, cutting, top and bottom coding, micro aggregation, adding noise, PRAM and rank exchange within a toolbox. For development of *OOSDCS*, Microsoft Visual Studio 2008 Professional Edition Evaluation Version software was used for the development environment. The programming language was chosen as C# as one of the languages supported by the software development environment. The reason for choosing C# is that it is an object-oriented platform-independent language. When objects in the *OOSDCS* Toolbox representing disclosure control methods and auxiliary functions are dragged and left in the workspace, relationships between them are created with arrows. Finally the system to set parameters belonging to the objects is implemented. The inputs are micro data and meta data files, while the output of the *OOSDCS* system are processed variables. Details related to the *OOSDCS* software are mentioned in the headings below.

4.2. OOSDCS Multiple Document Interface

OOSDCS uses a multiple document interface. By opening more than one workspace, it is possible to set up different systems in each workspace. Figure 1 displays the multiple document interface in *OOSDCS*. The toolbox on the left contains objects used to create the statistical disclosure control systems.

4.3. OOSDCS Menus

File Menu: Workspaces are opened by choosing the New Workspace link in the File menu. The created workspace automatically contains ordered names. In this version of *OOSDCS*, each workspace uses a palette to open disclosure control systems, with only outputs from these systems saved to disk instead of the systems themselves. When you wish to close the program, the Exit link on the File menu is used.

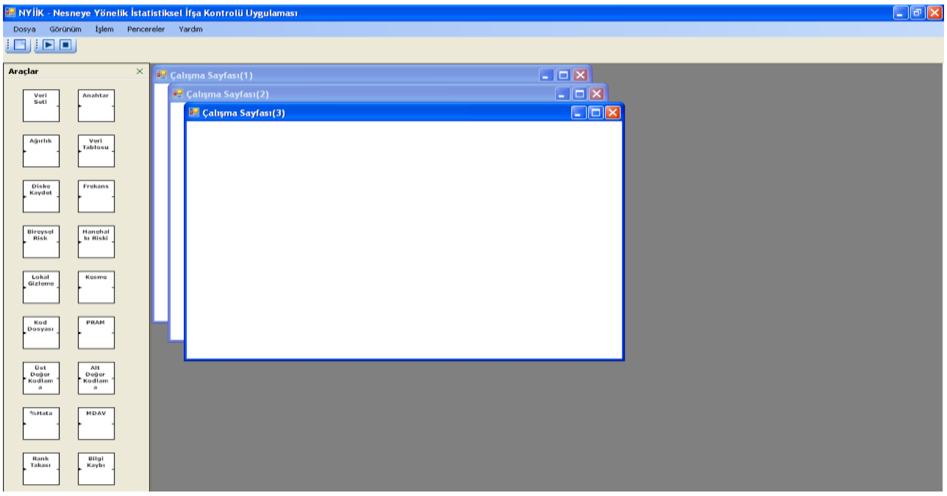


Figure 1. *OOSDCS* Multiple Document Interface

View Menu: By choosing the Toolbox link on the View menu, the closed toolbox is brought onscreen again.

Action Menu: To operate the *OOSDCS* systems from the workspace, the “Run” link is chosen from the Action Menu. To end a currently running system and reset the system, the “End Run” link is chosen from the Action menu.

Windows Menu: Open workspaces are listed in the Windows menu. To transition between workspaces, just choose the name of the requested workspace in this menu.

Help menu: To view the help file that accompanies *OOSDCS*, choose the HTML “Help” link from the Help menu.

4.4. OOSDCS Toolbox

As observed in Figure 1, the *OOSDCS* Toolbox contains eighteen tools. The aims and properties of some of these tools are explained under the headings below.

4.4.1. Data Set

The first component of the *OOSDCS* system setup each time is the data set. Data set tools are dragged and left on the workspace. When double-clicked with the mouse, the tools window in Figure 2 opens.

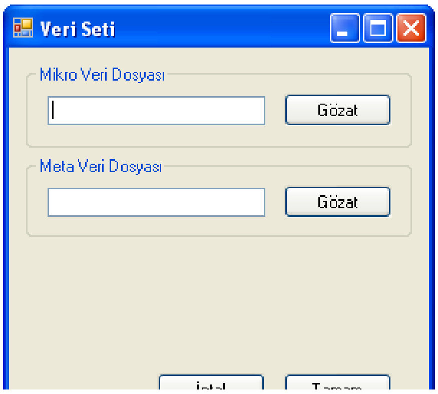


Figure 2. Settings Window for Data Set

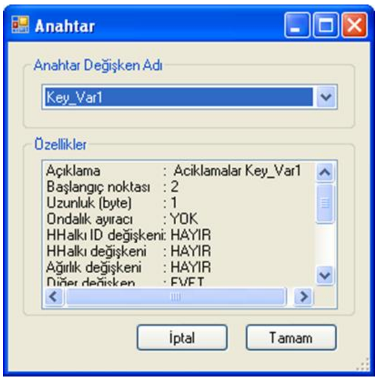
A data set comprises micro data and meta data (see Tables 2 and 3). Using the browse button, files in the file sequence may be chosen. During the choice process, micro data files are not uploaded to memory, only a path is determined. In meta data files the names and properties of key variables forming micro data are uploaded to memory when the OK button is clicked. The Data Set tool cannot form links toward inputs (left), but links may be created toward the output direction (right) with Key and Weight variables. The links between two tools are formed by pressing the left button of the mouse over the initial tool output and releasing the button when the input for the second tool is reached and is represented by an arrow. At the link points, the cursor is transformed into a plus sign to aid users.

4.4.2. Key

The key variables in the micro data file are represented by the Key tool on the workspace. For each of the key variables chosen for use in the *OOSDCS* system setup, a Key tool is added to the workspace. The settings window for the Key tool is shown in Figure 3. In the settings window, to choose the name of a Key variable there should be a link between each data set tool and the Key tool and it is necessary to have set up the data set tool. In the settings window for the Key tool, after the Key variable name is chosen from the list, the properties of the chosen key variable are listed in the Properties field. The Key tool can form links with the Data Set tool in the input direction (left). In the output direction (right), it may join with other key variables to form a data table or can be directly linked with disclosure control tools.

4.4.3. Weight

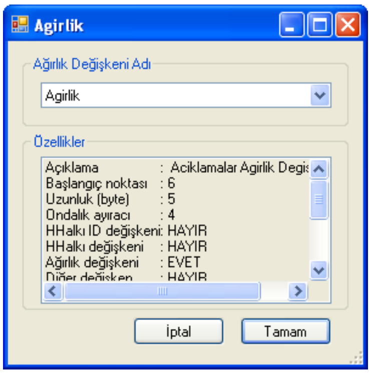
The weight of a variable represents the equivalence of a unit in a sample data set with units within an assemblage. If the data set is not a sample, all units have unit weight value. The weight variable is represented in the Weight tool in the *OOSDCS* Toolbox. The settings window for the Weight tool is shown in Figure 4. The Weight tool may link to the input of the Frequency tool to calculate frequency of key variables or data tables composed of key variables. The weight variable is not found in micro data sets compiled from full inventories. When calculating the frequency of records in these types of files, as the weight value is accepted as 1, there is a single link toward the input of the Frequency tool.



Anahtar Değişken Adı
Key_Var1

Özellikler
Açıklama : Aciklamalar Key_Var1
Başlangıç noktası : 2
Uzunluk (byte) : 1
Ondalık ayıracı : YOK
HHalkı ID değişkeni: HAYIR
HHalkı değişkeni : HAYIR
Ağırlık değişkeni : HAYIR
Diğer değişken : EVET

Figure 3. Settings for Key Variable



Ağırlık Değişkeni Adı
Agirlik

Özellikler
Açıklama : Aciklamalar Agirlik Degis
Başlangıç noktası : 6
Uzunluk (byte) : 5
Ondalık ayıracı : 4
HHalkı ID değişkeni: HAYIR
HHalkı değişkeni : HAYIR
Ağırlık değişkeni : EVET
Diğer değişken : HAYIR

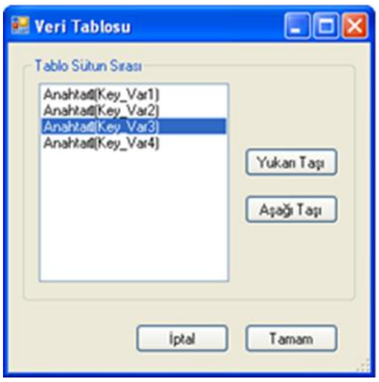
Figure 4. Settings Window for Weight Tool

4.4.4. Data Table

There is more than one input possible for the Data Table tool. The separate key variables represented in the Key tool unite under the Data Table object. For disclosure control methods applied for individual risk or household risk, as the correlation between many key variables are observed instead of a single key variable, firstly a data table is created of the key variables. With the condition that the contents have the same number of records, all of the tools in the *OOSDCS* toolbox may be linked with the data table. When information is written to file, the Save to Disk tool entry is generally a Data Table tool. The settings window for the Data Table tool is shown in Figure 5. Each of the tools linked to the Data Table inputs are inserted in the table as a column vector or column vector group. The first tool linked to the Data Table is represented by the first column. The column order may be changed by using the Move Up and Move Down buttons in the settingswindow. Organization of the Data Table does not affect disclosure control results, information locations are changed only in the file to be saved to disk.

4.4.5. Save to Disk

For the results produced by *OOSDCS* to be analyzed by the user, it is necessary to save to disk. The user determines which information is saved to disk in which order. In this version of *OOSDCS* when saving the results of a disclosure risk application to disk a new micro data file is not created, only records in memory are saved to disk. To change protected records and original records, tools in office programs may be used. The settings window for the Save to Disk tool is shown in Figure 6. In this window in the File Name field the ASCII text file name for the disk is written. The file path is determined by using the browse button. If information written in the file is to be in ordered columns, the Separate Table choice should be marked. If information is to be written consecutively like in the micro data file, the No Separation Mark choice should be marked. When ASCII text files are opened in Unicode or UTF text editors they are not displayed properly.

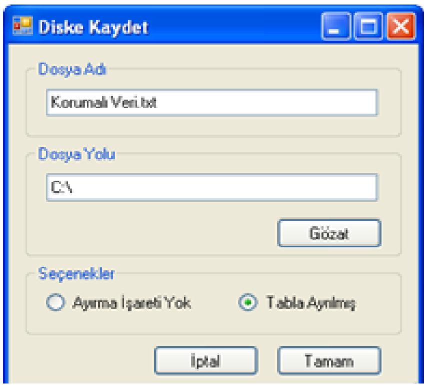


Tablo Sütun Sırası

Anahtar(Key_Var1)
Anahtar(Key_Var2)
Anahtar(Key_Var3)
Anahtar(Key_Var4)

Figure 5. Settings Window for Data Table Tool

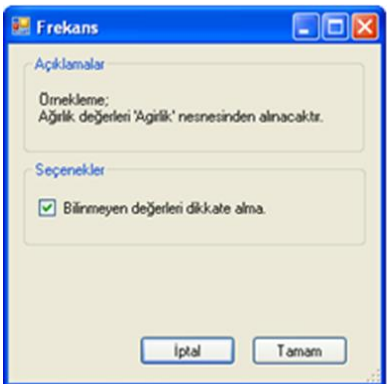
It is necessary to open them in WordPad or Notepad++ editors. For files saved to disk in formats other than ASCII, the characters may be represented by too many bits.



Dosya Adı: Korumalı Veri.txt

Dosya Yolu: C:\

Seçenekler: ○ Ayırma İşareti Yok ● Tabla Ayrılmış



Açıklamalar

Örnekleme;
Ağırlık değerleri 'Ağırlık' nesnesinden alınacaktır.

Seçenekler

☑ Bilinmeyen değerleri dikkate alma.

Figure 6. Setting Window for Save to Disk Tool

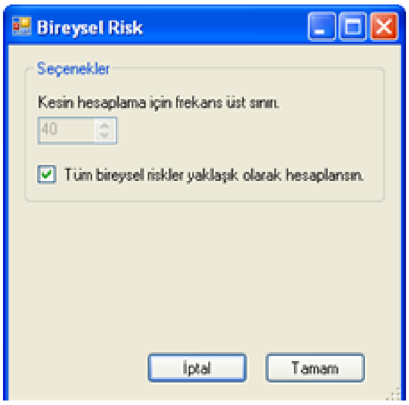
Figure 7. Settings for Frequency

4.4.6. Frequency

To calculate the individual risk or household risk of categorical data, it is necessary to know the values of the variable or the numbers in the sequences of variable clusters. There are two inputs; the first input in the Frequency tool is a Data Table or Key and the second is Weight. When full inventory data is used, the *OOSDCS* weight link is cancelled and the tool link number reduces to one. The output of the Frequency tool may link to Individual Risk or Household Risk tools or frequencies may be saved to disk. There is only one output from the Frequency tool; however this is combined with vectors in the single output. The settings menu for the Frequency tool is shown in Figure 7. In the explanation portion the user is informed about whether the data is a sample or full inventory depending on the presence of a Weight variable link. In the Options section, if there are unknown values within the data while calculating frequency, the user is asked to determine whether this value should be ignored or not.

4.4.7. Individual Risk

This is only used when calculating individual risk and frequencies. *OOSDCS* uses an approximate calculation formula by default, with definite calculations performed with the condition of determining the frequency upper limit. These settings are made in the Individual Risk tool settings window shown in Figure 8.



Bireysel Risk

Seçenekler

Kesin hesaplama için frekans üst sınırı.

40

☑ Tüm bireysel riskler yaklaşık olarak hesaplansın.

Figure 8. Settings for Individual Risk

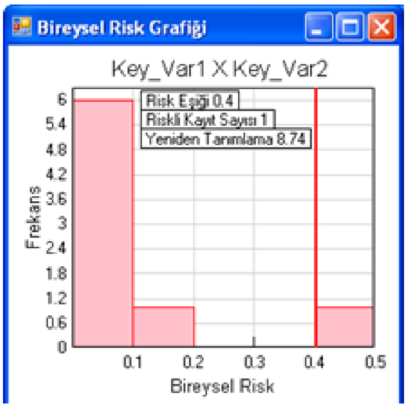


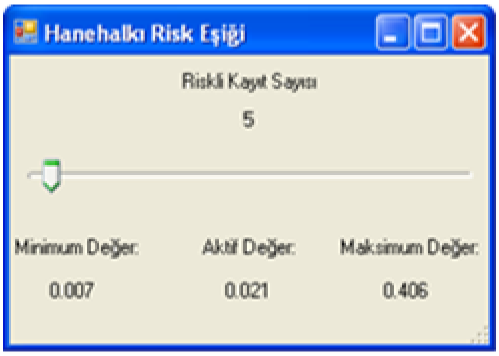
Figure 9. Graph Window for Individual Risk

As observed in Figure 9, after individual risk is calculated a column graph with risk on the horizontal axis and frequency on the vertical axis is presented for the user to determine risk threshold value. When the mouse is moved over this graph, risk threshold, number of records with risk and redefinition rates are actively viewed. By clicking on the left button of the mouse, the viewed values are placed in a frame and the risk threshold is determined. Links may be formed in the output direction from Individual Risk to Local Cloaking. When the graph window is closed, the system continues to operate from the last point.

4.4.8. Household Risk

In micro data for social research, data are generally compiled of households. While determining the risk for a household, a junction set of the individual risk carried by all individuals in a household is calculated. In households with risk above the determined household risk threshold, there are records the form a risk, in other words records increasing the mean risk. These records may later have local cloaking applied.

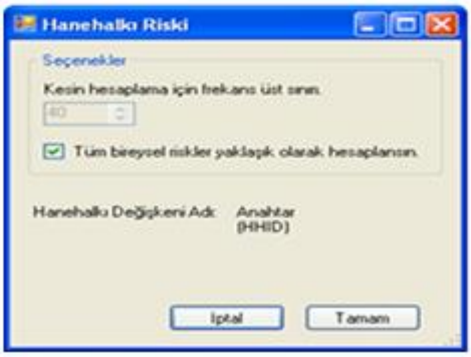
The individual risk value accepted as risky in a household may not be risky in another household with a different number of occupants. As determining records at risk on a graph is inefficient, the number of risk records equivalent to a certain interval of household risk threshold values are initially calculated by *OOSDCS* and the user then chooses one of these values. The household risk threshold value is chosen using a slider shown in Figure 10. There are two links to Household Risk inputs; the first is Frequency and the second is the Key tool representing the variable Household ID number. The settings window for Household Risk is shown in Figure 11. In this window there are selections relating to definite or approximate calculation of individual risk.



Riskli Kayıt Sayısı
5

Minimum Değer: 0.007
Aktif Değer: 0.021
Maksimum Değer: 0.406

Figure 10. Slider to determine Household Risk Threshold



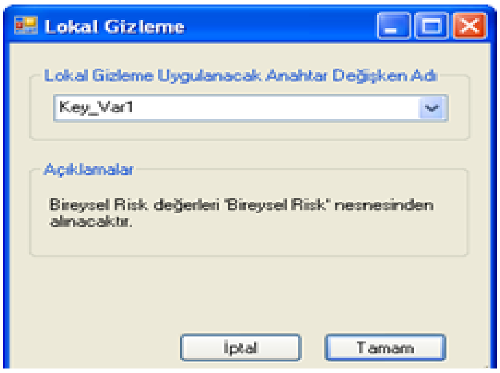
Seçenekler
Kesin hesaplama için frekans üst sınırı
40
☑ Tüm bireysel riskler yaklaşık olarak hesaplansın.

Hanehalkı Değişkeni Adı: Anahtar (HHID)

Figure 11. Settings Window for Household Risk

4.4.9. Local Cloaking

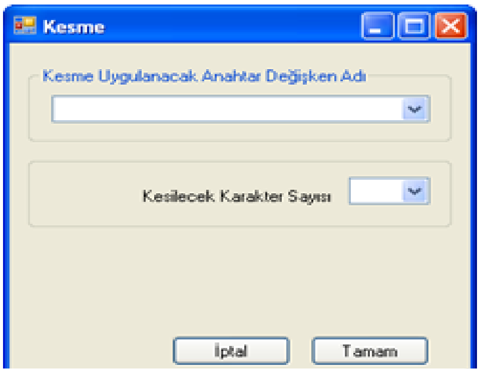
There are two input links to the Local Cloaking tool, one for Individual Risk and one for Key variable. Key variables above the Individual Risk threshold value have local cloaking applied and the cloaked values are represented with “ * ” character. To view the results, Local Cloaking is linked to the Save to Disk tool. The settings window for Local Cloaking is shown in Figure 12. The Key Variable Name for Local Cloaking is taken from the Key or Data Table linked to the Local Cloak input.



Lokal Gizleme Uygulanacak Anahtar Değişken Adı
Key_Var1

Açıklamalar
Bireysel Risk değerleri 'Bireysel Risk' nesnesinden alınacaktır.

Figure 12. Settings for Local Cloaking



Kesme Uygulanacak Anahtar Değişken Adı

Kesilecek Karakter Sayısı

Figure 13. Settings for Cutting

4.4.10. Cutting

Cutting is a global recoding method and may be applied to categorical variables with cuttable properties. As shown in the settings window for Cutting in Figure 13, for a categorical variable to be cuttable it must have more than one step and from right to left as step numbers decrease the represented area must widen. For example, statistical region classifications and international occupation and activity codes have this form.

4.4.11. Code File

The intervals of ordered categorical variables may be separated and each interval renamed. The intervals and new value corresponding to each interval is stated in the code file. The Code file format for *OOSDCS* is shown in Table 1. All values that a variable can have must be contained in the rows in the code file.

The settings window for the Code File tool, with input links from Key or Data Tables, is shown in Figure 14. After determining the Key Variable Name for Code File, the code file is chosen with the help of the browse button.

Table 1. *OOSDCS* Code File Format

NewValue<Tab>Value1<Tab>Value2 For example; for 5 year interval age groups: 0-4 0 4 5-9 5 9	**NewValue:** New value representing the remaining values between Value1 and Value2 **Value1:** Initial value of interval **Value2:** Final value of interval. Must be larger than initial value.

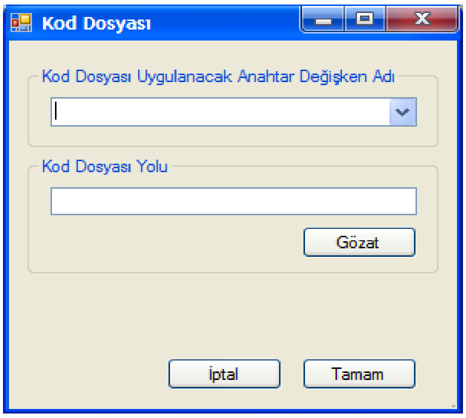


Figure 14. Settings for Code File Tool

Coding records reduces the individual risks carried; however it affects not only risky records but all records in the micro data file.

4.4.12. Top and Bottom Coding

When a numerical variable in a unit has a very high or very low value, the risk of redefinition of the unit increases. For example, an individual with very high age in survey data will be easily distinguished from other records. How the top and bottom limits are identified is linked to the strength of the definitive property of the variable and the intervals in which the values become more infrequent. Top and bottom coding is defined as two separate tools in the *OOSDCS* Toolbox for flexible use. These tools can be used consecutively on the same variable to complete both top and bottom coding.

Top and Bottom Coding tools may have input links from Key or Data Table tools. The windows used for Top and Bottom Coding tool settings are shown in Figure 15 and 16. In these windows the variable name to be coded and top and bottom limits of coding are determined.

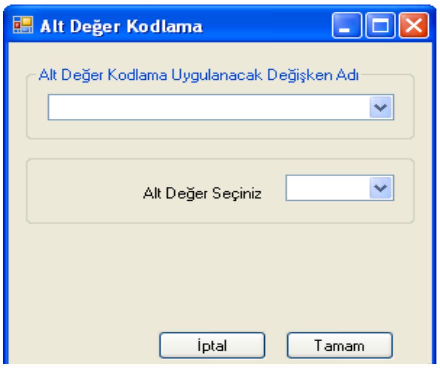


Figure 15. Settings Window for top

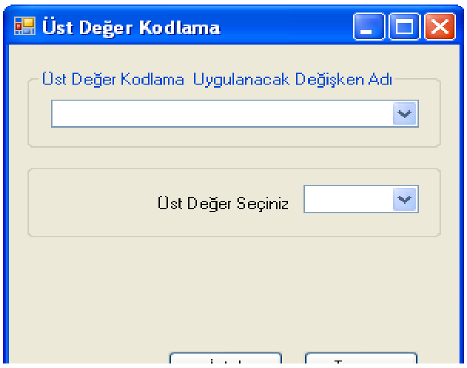


Figure 16. Settings Window bottom value coding

5. OOSDCS Local Cloaking Model

In this study, *OOSDCS* was tested using a sample micro data set and successful and unsuccessful aspects were assessed. In the example, the *Francdat synthetic data set* from the *sdcMicro* packet written in *R* language was used [11]. This data set was derived originally by Capobianchi, Polettini and Lucarelli for use in demonstration displays [12].

The Francdat micro data set contains 8 observational results for 8 variables. To provide an example of household risk, two variables of household number and unit number were added to the data set by the authors. Table 2 shows the Francdat data set format while Table 3 contains the meta and micro data file for the Francdata data set.

Table 2. Francdat Data Set Format

HHID	Household ID number 1 character length
UNITID	Unit number. 1 character length
Num1	Numerical vector. 4 character length. 2nd character is decimal separator
Key1	Key Variable 1. 1 character length
Num2	Numerical vector. 4 character length. 2nd character is decimal separator.
Key2	Key Variable 2. 1 character length
Key3	Key Variable 3. 1 character length
Key4	Key Variable 4. 1 character length
Num3	Numerical vector. 2 character length.
W	Weight vector. 5 character length. 4th character is decimal separator.

Table 3. Francdat Data Set Meta and Micro Data Files

META DATA FILE (francdatmeta.txt)	HHID -att 0,1,0 -typ 1,2 UNITID -att 1,1,0 Num1 -att 2,4,2 Key1 -att 6,1,0 -max 6 Num2 -att 7,4,2 Key2 -att 11,1,0 -max 3 Key3 -att 12,1,0 -max 5 Key4 -att 13,1,0 -max 5 Num3 -att 14,2,0 W -att 16,5,4 -typ 3,2

MICRO DATA FILE (francdat.txt)	110.3010.40251 4 18.0
	120.1210.2221122 45.5
	130.1810.80211 8 39.0
	141.9039.0031591 17.0
	251.0041.3031413541.0
	261.0041.4031114 8.0
	370.1060.01215 1 5.0
	380.1510.50251 5 92.0

When the *Key1, Key2, Key3* and *Key4* key variables and *W* weight are taken from the Francdat data set, the example *OOSDCS* mode in Figure 17 was created. According to the scenario in this model, first the chosen key variables need to be brought into a data table structure. Later the weight frequencies are identified from the weight variable values in data table records and then individual risk is calculated and the user is requested to determine risk threshold with the aid of the graph. According to the user's chosen risk threshold, local cloaking is applied to the current values of the Key 1 key variable and the new values are saved to disk. From the Action menu *"Run"* is chosen and the *OOSDCS* system operates. While the system is operating, after calculating individual risk the individual risk graph in Figure 18 appears on screen to determine individual risk threshold. With the aid of a mouse, risk threshold is determined on the graph. The risk threshold value is highlighted by a thick line on the graph. Records remaining on the right side of the risk threshold are accepted as risky and these are shown on the graph. Just as risk threshold may be chosen during the run time, it may be a previously determined value between 0 and 1. If a risk threshold of 1 is chosen, it means there is no cloaking of the micro data.

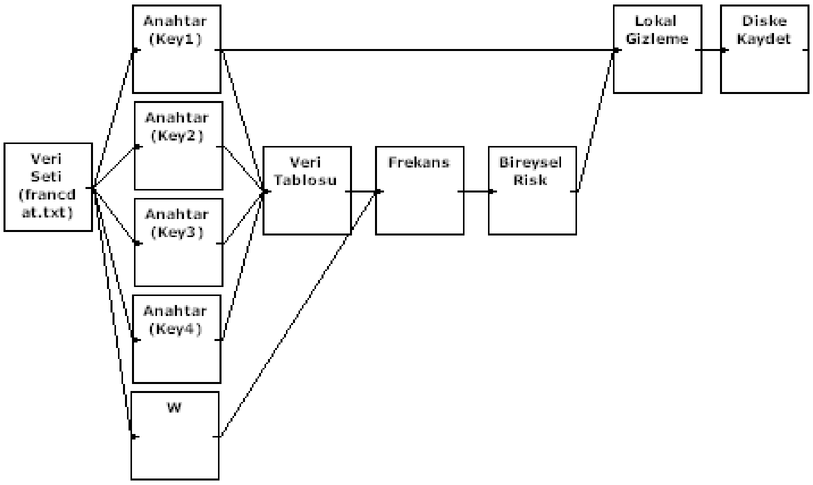


Figure 17. Sample *OOSDCS* Model set up to Apply Local Hide

On the individual risk graph in Figure 18, there is 1 risky record to the right of the chosen risk threshold value (0.398), and during local cloaking the value of this record will be changed to *. In the local cloaking model below, as the outputs of key variables are not linked, it runs simultaneously. For the Local Cloaking component to use the value belonging to the Key1 key variable, you must wait for the operation of the Individual Risk component to finish.

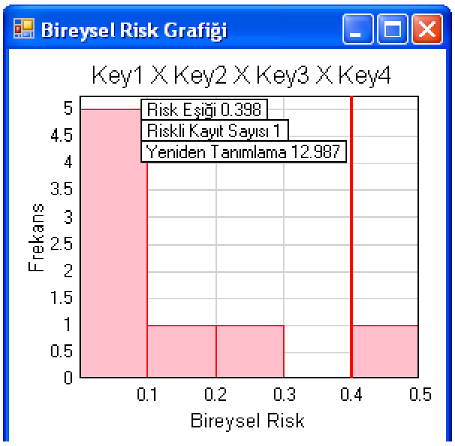


Figure 18. Individual Risk Graph and Determination of Risk Threshold on Graph

6. Results and Discussion

The demands of researchers and the public for high quality micro data have rapidly increased in the last number of years. Linked to regulations and ethical principles of data confidentiality, statistics offices, agencies and other organizations can provide masked data for use by researchers and the public. To produce high quality masked data, there is a need for statistical disclosure control software. Disclosure control methods applicable to micro data are generally local cloaking, global recoding, cutting, top and bottom coding, microaggregation, adding noise, PRAM and rank exchange. Software's to apply statistical confidentiality are available for free. The *μ-ARGUS* and *sdcMicro* software's are the most successful in this area. The *OOSDCS* developed within this study generally presents the statistical confidentiality methods offered by *μ-ARGUS* [13] and *sdcMicro* in hybrid, flexible and interactive form.

Additionally a different approach was considered for the user interface of *OOSDCS* and a toolbox was created. The objects in the toolbox represent confidentiality application methods and auxiliary functions and can be dragged onto the workspace and then relations can be created with arrows to create pipeline-structure *OOSDCS* systems. In addition to not requiring any knowledge of code, *OOSDCS* requires users to be aware of the inputs desired or needed for the methods applied and produced outputs. *OOSDCS* does not work as a black box, the workflow and process steps are completely designed by the user. All process steps in *OOSDCS* are modeled as a part of a system with pipeline structure, making it easy to report what occurs from the first stage to the last stage.

Though the initial creation of the *OOSDCS* system model may take time, once the command prompts for the developed system are entered, it does not just work once. The links between components in the system may be changed to create new systems or links may be created between two different *OOSDCS* systems in the workspace to create a single system. Due to this property, all stages of a statistical disclosure scenario can be built within the *OOSDCS* system. The operation of many tools at the same time in *OOSDCS* workspaces may make it difficult to visually analyze the systems. However, the general approach in statistical disclosure control is the use of only key variables with high descriptive properties not all variables and the choice of one or more statistical disclosure control methods.

OOSDCS offers the advantage of multi-threaded programming. While an *OOSDCS* system is operating, each link may produce a new program thread and operate independently of the others. Components (tools) with independent outputs may operate simultaneously. In this way, processing time is used more efficiently and productively than other software with classic procedural approaches and commands working in order.

References

[1] Quinto, W. and Singer, S. "Trade secrets : law and practice", Oxford University, Press New York , 2009.

[2] Elliot, M.J. and Dale, A.,"Scenarios of attack : the data intruder's perspective on statistical disclosure risk", Netherlands Official Statistical, Spring,1999, pp. 6-10.

[3] ASA, "Data Access and Personal Privacy: Appropriate Methods of Disclosure Control", American Statistical Association Notice, 2008.

[4] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz R., Longhurst, J.,. Schulte Nordholt, E., Seri, G., and de Wolf, P.P., "Handbook on Statistical Disclosure Control", 2009, vol 1.1, ESSnet SDC.

[5] Gouweleeuw, J., Kooiman, M., Willenborg, P., and Wolf, de P. P., "Post randomization for statistical disclosure control: Theory and implementation", Journal of Official Statistics, issue 14, 1998a, pp. 463-478.

[6] Gouweleeuw, J.M., P. Kooiman, L.C.R.J. Willenborg and P.P. de Wolf (1998b), The post randomisation method for protecting micropdata", Qüestiió, Quaderns d'Estadística i nvestigació Operativa, vol. 22 issue 1, 1998b, pp. 145 – 156.

[7] Warner, S.L,"Randomized Response; a survey technique for eliminating evasive answer bias", Journal of the American Statistical Association, vol. 57, 1965, pp. 622 - 627.

[8] Hundepool, A., de Wetering, A.V., Ramaswamy, R., Franconi, L., Capobianchi, A., De Wolf, P.P., Domingo J. F., Torra, V., Brand, R., and Giessing, S., "μ-ARGUS version 4.2 Software and User's Manual", Statistics Netherlands, Voorburg NL, 2008.

[9] Templ, M., "Statistical Disclosure Control for Microdata Using the R-Package sdcMicro", Transactions on Data Privacy, vol.1-2, 2008, pp. 67 – 85.

[10] Manning, A. M., and Haglin, D. J., "A new algorithm for finding minimal sample uniques for use in statistical disclosure assessment", IEEE International Conference on Data Mining (ICDM05), Nov. 2005, pp 290-297.

[11] Templ, M., "sdcMicro: Statistical Disclosure Control methods for the generation of public- and scientific-use files", 2009. http://cran.r-project.org/web/packages/sdcMicro

[12] Capobianchi, A., Polettini, S., and Lucarelli, M., "Strategy for the implementation of individual risk methodology into μ-ARGUS", Technical report, Report for the CASC project. No: 1.2-D1, 2001.

[13] de Wolf , P.P., Hundepool, A., Giessing, S., Salazar , J.J., Castro, J., "μ-argus version 4.1 software and user's manual", Argus Open Source-project, Statistics Netherland, P.O. Box 24500, 2014.