University of Texas at El Paso
ScholarWorks@UTEP

Departmental Technical Reports (CS) | Computer Science

3-2013

Data Anonymization that Leads to the Most Accurate Estimates of Statistical Characteristics: Fuzzy-Motivated Approach

G. Xiang
Applied Biomathematics, gxiang@sigmaxi.net

S. Ferson
Applied Biomathematics

L. Ginzburg
Applied Biomathematics

L. Longpre
The University of Texas at El Paso, longpre@utep.edu

E. Mayorga
The University of Texas at El Paso

Follow this and additional works at: https://scholarworks.utep.edu/cs_techrep

 Part of the Computer Sciences Commons

See next page for additional authors

Technical Report: UTEP-CS-13-12a

To appear in *Proceedings of the joint World Congress of the International Fuzzy Systems Association and Annual Conference of the North American Fuzzy Information Processing Society IFSA/NAFIPS'2013*, Edmonton, Canada, June 24-28, 2013.

Recommended Citation

Xiang, G.; Ferson, S.; Ginzburg, L.; Longpre, L.; Mayorga, E.; and Kosheleva, O., "Data Anonymization that Leads to the Most Accurate Estimates of Statistical Characteristics: Fuzzy-Motivated Approach" (2013). *Departmental Technical Reports (CS)*. 741.
https://scholarworks.utep.edu/cs_techrep/741

This Article is brought to you for free and open access by the Computer Science at ScholarWorks@UTEP. It has been accepted for inclusion in Departmental Technical Reports (CS) by an authorized administrator of ScholarWorks@UTEP. For more information, please contact lweber@utep.edu.

Authors
G. Xiang, S. Ferson, L. Ginzburg, L. Longpre, E. Mayorga, and O. Kosheleva

This article is available at ScholarWorks@UTEP: https://scholarworks.utep.edu/cs_techrep/741

Data Anonymization that Leads to the Most Accurate Estimates of Statistical Characteristics: Fuzzy-Motivated Approach

G. Xiang, S. Ferson, L. Ginzburg
Applied Biomathematics
100 North Country Rd.
Setauket, NY 11733, USA
contact email gxiang@sigmaxi.net

L. Longpré, E. Mayorga, O. Kosheleva
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
contact email olgak@utep.edu

***Abstract*—To preserve privacy, the original data points (with exact values) are replaced by boxes containing each (inaccessible) data point. This privacy-motivated uncertainty leads to uncertainty in the statistical characteristics computed based on this data. In a previous paper, we described how to minimize this uncertainty under the assumption that we use the same standard statistical estimates for the desired characteristics. In this paper, we show that we can further decrease the resulting uncertainty if we allow fuzzy-motivated *weighted* estimates, and we explain how to optimally select the corresponding weights.**

I. Formulation of the Problem

Need to preserve privacy. In many practical applications, e.g., in medicine and in education, to better serve customers, it is important to know as much as possible about the potential customers. Customers are often reluctant to share information, since this information can be potentially used against them. For example, age can be used by companies to (unlawfully) discriminate against older job applicants. It is thus important to preserve privacy when storing customer data; see, e.g., [6].

How to preserve privacy: k-anonymity and ℓ-diversity. To maintain privacy, we divide the space of all possible combinations of values $x = (x_1, \ldots, x_n)$ into boxes

$$B = [\widetilde{x}_1 - \Delta_1(x), \widetilde{x}_1 + \Delta_1(x)] \times \ldots \times [\widetilde{x}_n - \Delta_n(x), \widetilde{x}_n + \Delta_n(x)]. \quad (1)$$

For each record, instead of storing the actual values x_i, we only store the label of the box B containing x.

To avoid further loss of privacy, it is important to make sure that location in a box does not identify a person. This is usually achieved by requiring that for some fixed integer k, each box contains at least k records.

It is also not good if all records within a box have the same value of an i-th quantity x_i. It is thus required that for some integer ℓ, each box should contain at least ℓ different values of each x_i; see, e.g., [1].

Statistical data processing. Based on the available data points $x^{(p)} = \left(x_1^{(p)}, \ldots, x_n^{(p)}\right)$ $(1 \le p \le N)$, we need to estimate averages E_i, variances $V_i = \sigma_i^2$, covariances C_{ij}, correlations ρ_{ij}, and other statistical characteristics. The means are usually estimated as follows:

$$E_i = \frac{1}{N} \cdot \sum_{p=1}^{N} x_i^{(p)}, \quad E_j = \frac{1}{N} \cdot \sum_{p=1}^{N} x_j^{(p)}. \quad (2)$$

The covariance is usually estimated as:

$$C_{ij} = \frac{1}{N} \cdot \sum_{p=1}^{N} \left(x_i^{(p)} - E_i\right) \cdot \left(x_j^{(p)} - E_j\right). \quad (3)$$

The variance is usually estimated by a formula

$$V_i = \frac{1}{N-1} \cdot \sum_{p=1}^{N} \left(x_i^{(p)} - E_i\right)^2 \quad (4)$$

or, sometimes,

$$V_i = \frac{1}{N} \cdot \sum_{p=1}^{N} \left(x_i^{(p)} - E_i\right)^2, \quad (5)$$

and the correlation is estimated as

$$\rho_{ij} = \frac{C_{ij}}{\sigma_i \cdot \sigma_j}. \quad (6)$$

Comment. We are interested in large databases, in which the number N of records is large. For large N, the difference between the usual un-biased estimate for V_i (with $N-1$ in the denominator) and the estimate with N is negligible. To simplify computations, in this paper, by V_i and σ_i, we will mean the versions corresponding to N; our results can be easily be reformulated for the un-biased estimates, which in our terms take the form $V_i \cdot \frac{N-1}{N}$ and $\sigma_i \cdot \frac{\sqrt{N-1}}{\sqrt{N}}$.

In statistical data processing, privacy leads to uncertainty. To maintain privacy, we replace each numerical value $x_i^{(p)}$ with the corresponding interval. Different values from these intervals lead, in general, to different values of the resulting

statistical characteristics. Hence, for each characteristic, we get a whole interval of possible values.

If this interval is too wide, the resulting range is useless (e.g., for correlation, the interval $[-1,1]$ is useless). It is therefore desirable to select, among all possible subdivisions into boxes which preserve k-anonymity (and ℓ-diversity), the one which leads to the narrowest intervals for the desired statistical characteristic.

What we do in this paper. In Section 2, following [7], we describe how this problem is solved now. Please note that because our objective is to generalize these formulas to the weighted case, the notations that we use in Section 2 are slightly different from the notations from [7].

Then, in Section 3, we explain how fuzzy-motivated ideas can improve the corresponding estimates.

II. How This Problem Is Solved Now

Estimating accuracy caused by privacy-based subdivision into boxes: case of k-anonymity. To minimize uncertainty, we select the smallest boxes. Hence, each box B should have exactly k records.

For each combination of values $x_i^{(p)}$ from the corresponding intervals $\left[\widetilde{x}_i^{(p)} - \Delta_i^{(p)}, \widetilde{x}_i^{(p)} + \Delta_i^{(p)}\right]$, we get:

$$C\left(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(N)}\right) =$$

$$C\left(\widetilde{x}_1^{(1)} + \Delta x_1^{(1)}, \widetilde{x}_2^{(1)} + \Delta x_2^{(1)}, \ldots, \widetilde{x}_n^{(N)} + \Delta x_n^{(N)}\right), \quad (7)$$

where each difference $\Delta x_k^{(p)} \stackrel{\text{def}}{=} x_k^{(p)} - \widetilde{x}_k^{(p)}$ satisfies the inequality $\left|\Delta x_i^{(p)}\right| \le \Delta_i^{(p)}$. When we have many records, boxes are small, so we can use a linear approximation:

$$C = \widetilde{C} + \sum_{p=1}^{N} \sum_{i=1}^{n} \frac{\partial C}{\partial x_i} \cdot \Delta x_i^{(p)}, \quad (8)$$

where $\widetilde{C} \stackrel{\text{def}}{=} C\left(\widetilde{x}_1^{(1)}, \widetilde{x}_2^{(1)}, \ldots, \widetilde{x}_n^{(N)}\right)$. The range of this linear expression is $\left[\widetilde{C} - \Delta, \widetilde{C} + \Delta\right]$, where

$$\Delta \stackrel{\text{def}}{=} \sum_{p=1}^{N} \sum_{i=1}^{n} \left|\frac{\partial C}{\partial x_i}\right| \cdot \Delta_i^{(p)} =$$

$$k \cdot \sum_{B} \sum_{x \in B} \sum_{i=1}^{n} \left|\frac{\partial C}{\partial x_i}\right| \cdot \Delta_i(x). \quad (9)$$

Expressions for the corresponding partial derivatives. The estimate for the accuracy Δ is described in terms of partial derivatives $\dfrac{\partial C}{\partial x_i}$ of the statistical characteristic C. For the mean E_i, the derivative is equal to

$$\frac{\partial E_i}{\partial x_i} = \frac{1}{N}. \quad (10)$$

For the variance V_i, we have

$$\frac{\partial V_i}{\partial x_i} = \frac{2 \cdot (x_i - E_i)}{N}. \quad (11)$$

Therefore, for $\sigma_i = \sqrt{V_i}$, we get

$$\frac{\partial \sigma_i}{\partial x_i} = \frac{x_i - E_x}{N \cdot \sigma_i}. \quad (12)$$

For the covariance C_{ij}, we have

$$\frac{\partial C_{ij}}{\partial x_i} = \frac{x_j - E_j}{N}. \quad (13)$$

For the correlation ρ_{ij}, we have:

$$\frac{\partial \rho_{ij}}{\partial x_i} = \frac{1}{N} \cdot \frac{(x_j - E_j) - \dfrac{C_{ij}}{\sigma_i^2} \cdot (x_i - E_i)}{\sigma_i \cdot \sigma_j}. \quad (14)$$

For all these characteristics C, the derivative takes the form

$$\frac{\partial C}{\partial x_i} = \frac{1}{N} \cdot b_i(x) \quad (15)$$

for some expression $b_i(x)$.

Towards optimal subdivision into boxes. The overall expression for Δ is a sum of terms corresponding to different points. So, to minimize Δ, we must, for each point, minimize the corresponding term

$$\sum_{i=1}^{n} \left|\frac{\partial C}{\partial x_i}\right| \cdot \Delta_i(x). \quad (16)$$

Because of the relation between the partial derivatives and $b_i(x)$, this minimization is equivalent to minimizing the term $\sum\limits_{i=1}^{n} a_i(x) \cdot \Delta_i(x)$, where we denoted $a_i(x) \stackrel{\text{def}}{=} |b_i(x)|$.

The only constraint on the values $\Delta_i(x)$ is that the corresponding box should contain exactly k different points. The number of points can be obtained by multiplying the data density $\rho(x)$ by the box volume $\prod\limits_{i=1}^{n} (2\Delta_i(x))$. The data density can be estimated based on the data. So, we minimize the expression

$$\sum_{i=1}^{n} a_i(x) \cdot \Delta_i(x) \quad (17)$$

under the constraint

$$\rho(x) \cdot 2^n \cdot \prod_{i=1}^{n} \Delta_i(x) = k. \quad (18)$$

(Asymptotically) optimal subdivision into boxes (case of k-anonymity). The Lagrange multiplier technique leads to

$$\Delta_i(x) = \frac{c(x)}{a_i(x)}, \quad (19)$$

for some $c(x)$. From the constraint (18), we get

$$c(x) = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)} \cdot \prod_{j=1}^{n} a_j(x)}. \quad (20)$$

This means that around each point x, we need to select the box with half-widths

$$\Delta_i(x) = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)}} \cdot \frac{\sqrt[n]{\prod_{j=1}^{n} a_j(x)}}{a_i(x)}. \tag{21}$$

The resulting accuracy is equal to

$$\Delta = \frac{n}{N} \cdot \sum_x c(x), \tag{22}$$

where the sum is taken over all N data points x.

We need to dismiss rare points. In many practical situations, we have rare points, for which the smallest box containing k of them is huge. Such a big-size box will contribute a large amount of uncertainty to Δ; so we should dismiss such rare points.

If we select a subset $S \subset \{1, 2, \ldots, N\}$ of the set of N original points, then the privacy-related uncertainty reduces to

$$\frac{n}{\#S} \cdot \sum_{x \in S} c(x), \tag{23}$$

where $\#(S)$ denote the number of points in the set S. The statistical accuracy reduces to

$$\frac{A}{\sqrt{\#(S)}} \tag{24}$$

(see, e.g., [5]). Minimizing the sum

$$\frac{n}{\#(S)} \cdot \sum_{x \in S} c(x) + \frac{A}{\sqrt{\#(S)}} \tag{25}$$

leads to selecting all x with $c(x) \le c_0$, where c_0 minimizes the sum

$$\frac{n}{\#\{x : c(x) \le c_0\}} \cdot \sum_{x : c(x) \le c_0} c(x) + \frac{A}{\sqrt{\#\{x : c(x) \le c_0\}}}. \tag{26}$$

Examples. For estimating the mean E_i, we have $a_i(x) = \text{const}$ and thus,

$$c(x) = \text{const} \cdot \frac{1}{\sqrt[n]{\rho(x)}}. \tag{27}$$

In this case, $c(x)$ is a decreasing function of density. So, dismissing points with $c(x) > c_0$ is equivalent to dismissing all the points with $\rho(x) < \rho_0$ (for some ρ_0).

For computing covariance C_{ij}, the derivative is proportional to $x_i - E_i$. Thus, the values $a_i(x)$ are proportional to $|x_i - E_i|$. So, the upper threshold c_0 on $c(x)$ is equivalent to the lower threshold on the ratio

$$\frac{\rho(x)}{|x_i - E_i| \cdot |x_j - E_j|}. \tag{28}$$

Hence, we can also use points x with small $\rho(x)$, provided that if x_i or x_j is close to the corresponding mean. Using extra points x improves accuracy.

How to also take into account ℓ-diversity. Up to now, we only took into account the k-anonymity requirement. We also need to take into account that within each box, for each variable x_i, there are $\ge \ell$ different values of x_i. To formalize this requirement, we first need to describe what "different" means.

Usually, for each variable i, different means that

$$|x_i - x_i'| \ge \varepsilon_i \tag{29}$$

for some threshold ε_i. Thus, ℓ different values means that $2\Delta_i(x) \ge \ell \cdot \varepsilon_i$. So, the problem is to find $\Delta_i(x)$ such that

$$\sum_{i=1}^{n} a_i(x) \cdot \Delta_i(x) \to \min \tag{30}$$

under the constraints

$$\prod_{i=1}^{n} \Delta_i(x) \ge \frac{k}{2^n \cdot \rho(x)} \tag{31}$$

and

$$2\Delta_i(x) \ge \ell \cdot \varepsilon_i \tag{32}$$

for all i.

According to [7], the solution to this optimization problem is as follows: around each point x, we first compute the values

$$\Delta_i(x) = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)}} \cdot \frac{\sqrt[n]{\prod_{j=1}^{n} a_j(x)}}{a_i(x)}. \tag{33}$$

If $2\Delta_i(x) \ge \ell \cdot \varepsilon_i$ for all i, we select $\Delta_i(x)$. Otherwise, we sort the quantities by $a_i(x) \cdot \varepsilon_i$:

$$a_1(x) \cdot \varepsilon_1 \ge a_2(x) \cdot \varepsilon_2 \ge \ldots \ge a_n(x) \cdot \varepsilon_n. \tag{34}$$

Then, for each t from 1 to n, we compute

$$c_t = \frac{1}{2} \cdot \left(\frac{k \cdot \prod_{i=t+1}^{n} a_i(x)}{\rho(x) \cdot \ell^t \cdot \prod_{i=1}^{t} \varepsilon_i} \right)^{1/(n-t)}. \tag{35}$$

For each t, if $\dfrac{2c_t}{\ell} \ge a_{t+1}(x) \cdot \varepsilon_{t+1}$, we compute

$$\Delta(t) \stackrel{\text{def}}{=} \frac{1}{2} \cdot \ell \cdot \sum_{i=1}^{t} a_i(x) \cdot \varepsilon_i + (n-t) \cdot c_t. \tag{36}$$

We select t for which $\Delta(t)$ is the smallest, and take:

- $\Delta_i(x) = \dfrac{1}{2} \cdot \ell \cdot \varepsilon_i$ for $i \le t$, and
- $\Delta_i(x) = \dfrac{c_t}{a_i(x)}$ for $i > t$.

Comment. The computation time of this algorithm is quadratic in n. This is OK, since the number n of different characteristics is usually reasonably small. What is important is that the algorithm is still linear-time in terms of the number of records N.

III. Fuzzy-Motivated Idea

Main idea. In [7], to improve the accuracy of the resulting estimate, we propose to ignore some data points while keeping other data points. In other words, we propose a crisp separation between data points that we keep and data points that we ignore. Fuzzy logic has taught us that in many cases, it is beneficial to replace such a crisp separation with a "fuzzy" one in which, instead of ignoring or keeping a data point, we take a data point with a certain degree; see, e.g., [2], [4], [8].

Implementing the idea. Specifically, instead of using the above formula for computing the statistical characteristics, in which all data points are treated equally, we assign a weight $w(x) \geq 0$ to each data point so that $\sum w(x) = 1$, and use the weighted estimates for all the statistical characteristics:

$$E_i = \sum_x w(x) \cdot x_i, \quad \sigma_i^2 = \sum_x w(x) \cdot (x_i - E_i)^2, \qquad (37)$$

$$C_{ij} = \sum_x w(x) \cdot (x_i - E_i) \cdot (x_j - E_j), \quad \rho_{ij} = \frac{C_{ij}}{\sigma_i \cdot \sigma_j}. \qquad (38)$$

Optimization problem. Our objective is to find the weights $w(x)$ for which the resulting uncertainty is the smallest possible. Similarly to the crisp case, this uncertainty consists of two parts: the part coming from the privacy-motivated uncertainty and the part coming from the fact that the size is finite.

One can check that for privacy-motivated uncertainty, the corresponding derivatives $\dfrac{\partial C}{\partial x_i}$ are proportional to the weight $w(x)$. For each box, we thus face the exact same optimization problem for finding the best sizes $\Delta_i(x)$ of the corresponding privacy-related box. As a result, for the overall privacy-motivated uncertainty, we get the expression $n \cdot \sum\limits_x w(x) \cdot c(x)$.

For the statistical part: if we simply estimate the variance of the estimate for the mean $E_i = \sum w(x) \cdot x_i$, then, due to the fact that the variance of the sum of independent variables is equal to the sum of the variances, we conclude that the variance of this estimate is proportional to $\sum w^2(x)$; see, e.g., [5]. Thus, the standard deviation of this estimate is proportional to

$$\sqrt{\sum_x w^2(x)}. \qquad (39)$$

For the traditional equal-weight estimate, when

$$w(x) = \frac{1}{\#(S)} \qquad (40)$$

for all x, the proportionality coefficient becomes equal to the expression

$$\frac{1}{\sqrt{\#(S)}} \qquad (41)$$

that we used in Section 2.

One can check that, similarly, estimates for the accuracy of other statistical characteristics can be obtained from the estimates provided in Section 2 by replacing $\dfrac{1}{\#(S)}$ with $\sqrt{\sum\limits_x w^2(x)}$, i.e., this part is equal to

$$A \cdot \sqrt{\sum_x w^2(x)}. \qquad (42)$$

Thus, to minimize the overall inaccuracy, we need to minimize the following sum:

$$n \cdot \sum_x w(x) \cdot c(x) + A \cdot \sqrt{\sum_x w^2(x)} \qquad (43)$$

under the constraints $\sum\limits_x w(x) = 1$ and $w(x) \geq 0$.

Solving the resulting optimization problem: general idea. By applying the Lagrange multiplier method to the above constraint optimization problem, we can reduce this problem to the following unconstrained optimization problem:

$$n \cdot \sum_x w(x) \cdot c(x) + A \cdot \sqrt{\sum_x w^2(x)} \; - \; \lambda \cdot \left(\sum_x w(x) - 1 \right) \to \min, \qquad (44)$$

for an appropriate Lagrange multiplier λ. Differentiating this objective function with respect to $w(x)$ and equating the derivative to 0, we conclude that

$$n \cdot c(x) + \frac{A \cdot w(x)}{\sqrt{\sum\limits_y w^2(y)}} - \lambda = 0, \qquad (45)$$

i.e., that

$$w(x) = \frac{1}{A} \cdot (\lambda - n \cdot c(x)) \cdot \sqrt{\sum_y w^2(y)}. \qquad (46)$$

To be more precise, since we require that $w(x) \geq 0$, this formula only holds when $n \cdot c(x) \leq \lambda$; when $n \cdot c(x) > \lambda$, we should get $w(x) = 0$.

Towards computing the auxiliary parameter λ. How can we find λ? By squaring both sides of this formula, we get

$$w^2(x) = \frac{1}{A^2} \cdot (\lambda - n \cdot c(x))^2 \cdot \sum_y w^2(y). \qquad (47)$$

By adding left- and right-hand sides corresponding to different x, we get

$$\sum_x w^2(x) = \frac{1}{A^2} \cdot \left(\sum_x (\lambda - n \cdot c(x))^2 \right) \cdot \sum_y w^2(y). \qquad (48)$$

Dividing both sides of this equality by $\sum\limits_x w^2(x) = \sum\limits_y w^2(y)$, we conclude that

$$1 = \frac{1}{A^2} \cdot \sum_x (\lambda - n \cdot c(x))^2, \qquad (49)$$

i.e., that

$$\sum_x (\lambda - n \cdot c(x))^2 - A^2 = 0. \quad (50)$$

This is a quadratic equation in terms of λ, namely:

$$\widetilde{N} \cdot \lambda^2 - 2\lambda \cdot n \cdot \sum_x c(x) + n^2 \cdot \sum_x c^2(x) - A^2 = 0, \quad (51)$$

where $\widetilde{N}$ is the total number of points that we did not dismiss, i.e., for which $n \cdot c(x) < \lambda$, and the sums are taken over all such points.

From this quadratic equation, we can find λ. Thus, we naturally arrive at the following iterative algorithm for computing λ.

Iterative algorithm for computing the auxiliary parameter λ. The goal of this algorithm is to find the threshold value λ, so that points x for which $n \cdot c(x) \geq \lambda$ will be dismissed from our estimates (i.e., we would have $w(x) = 0$ for such points).

In the beginning, we do not have any reason to dismiss any values, so we start with the first approximation λ_0.

On each iteration k, we start with the value λ_{k-1} obtained on the previous iteration, and compute the next approximation λ_k as follows.

- First, we compute the total numbers $\widetilde{N}$ of points x for which $n \cdot c(x) < \lambda_{k-1}$.
- Then, we compute the sums $\sum_x c(x)$ and $\sum_x c^2(x)$ over all such points.
- Based on these values, we solve the quadratic equation (51) and find the next approximation λ_k.

We stop iterations when the process converges, i.e., when

$$\lambda_k = \lambda_{k-1}. \quad (52)$$

Towards computing $w(x)$. We know, from the formula (46), that for those points for which $n \cdot c(x) < \lambda$, we have

$$w(x) = K \cdot (\lambda - c(x)), \quad (53)$$

for some constant K. To find K, we can use the fact that $\sum_x w(x) = 1$. Substituting the expression (53) into this constraint, we conclude that

$$1 = K \cdot \left(\widetilde{N} \cdot \lambda - \sum_x c(x) \right). \quad (54)$$

Since we have already computed the values $\widetilde{N}$, λ, and $\sum_x c(x)$ when we computed λ, we can thus compute K.

So, we arrive at the following formula for computing the desired weights.

Formula for computing the optimal weights $w(x)$. By running the above iterative algorithm, we have computed the auxiliary value λ. In the process of computing λ, we have computed the values $\widetilde{N}$ and $\sum_x c(x)$, where the sum is taken over all the points x for which $n \cdot c(x) < \lambda$.

Now, we compute

$$K = \frac{1}{\widetilde{N} \cdot \lambda - \sum_x c(x)}. \quad (55)$$

The optimal weights can now be computed as follows:

- when $n \cdot c(x) \geq \lambda$, the optimal weight is $w(x) = 0$;
- when $n \cdot c(x) < \lambda$, the optimal weight is equal to

$$w(x) = K \cdot (\lambda - c(x)). \quad (56)$$

Comment. As expected, the larger the uncertainty contribution $c(x)$ from a point, the smaller the weight with which we take this point. When this contribution is large enough (i.e., larger than the threshold determined by the auxiliary parameter λ), we completely ignore such points.

IV. Boxes Appropriate for Several Different Characteristics

What we provided before. In the previous sections, we described how, for each statistical characteristic C, we can find the boxes (i.e., data anonymization) that leads to the most accurate estimate of this selected characteristic.

Remaining problem. In practice, we may need to compute the values of different statistical characteristics. The problem is that optimal boxes corresponding to different characteristics C are, in general, different.

For example, boxes that lead to most accurate estimates $\widetilde{E}$ of mean E may lead to very inaccurate estimates $\widetilde{C}_{ij}$ of correlation C_{ij}, and vice versa.

Towards a possible solution to this problem. Based on the previous experience, we know how many times users were looking for values of different statistical characteristics; in other words, we know the probabilities $p_C \geq 0$ $\left(\sum_C p_C = 1 \right)$ of looking for different characteristics C.

We also know what accuracy Δ_0^C is desirable for estimating each characteristic C. For example, we may fix the same relative error for all estimates, and take, e.g., $\Delta_0^C = 0.1 \cdot \widetilde{C}$ if this relative error is 10%. Then, for each characteristic C, the accuracy of estimating this characteristic is better gauged not by the absolute accuracy Δ^C but rather by the ratio

$$q_C \stackrel{\text{def}}{=} \frac{\Delta^C}{\Delta_0^C} \quad (57)$$

describing how close we are to the desired accuracy.

In this situation, a reasonable idea is to minimize *average* quality

$$q \stackrel{\text{def}}{=} \sum_C p_C \cdot q_C. \quad (58)$$

Towards an algorithm. How can we solve the corresponding optimization problem? The objective function q has the form

$$q = \sum_C \frac{p_C}{\Delta_0^C} \cdot \Delta^C, \quad (59)$$

i.e., the for

$$q = \sum_C \frac{p_C}{\Delta_0^C} \cdot \sum_{p=1}^{N} \sum_{i=1}^{n} \left| \frac{\partial C}{\partial x_i} \right| \cdot \Delta_i^{(p)}. \tag{60}$$

By changing the order of summation, we get an equivalent formula

$$q = \sum_{p=1}^{N} \sum_{i=1}^{n} \left(\sum_C \frac{p_C}{\Delta_0^C} \cdot \left| \frac{\partial C}{\partial x_i} \right| \right) \cdot \Delta_i^{(p)}. \tag{61}$$

This optimization problem is similar to the optimization problem corresponding to the case of a single statistical characteristic C, with the only difference that instead of the original partial derivatives $\left| \frac{\partial C}{\partial x_i} \right|$, we use a weighted combination

$$\sum_C \frac{p_C}{\Delta_0^C} \cdot \left| \frac{\partial C}{\partial x_i} \right| \tag{62}$$

of these derivatives.

In terms of the coefficients $a_i(x)$ introduced in Section 2, this means that instead of using the values $a_i^C(x)$ corresponding to an individual characteristic C, we must use a linear combination of these values:

$$a_i(x) = \sum_C \frac{p_C}{\Delta_0^C} \cdot a_i^C(x). \tag{63}$$

Resulting algorithm. Use the same algorithm(s) as in Sections 2 and 3, except that instead of the values a_i^C corresponding to an individual statistical characteristic C, we should use the values (63).

ACKNOWLEDGMENT

Support for this project was provided by the National Institutes of Health (NIH), through a Small Business Innovation Research grant (award number 1R43TR000173-01) to Applied Biomathematics, but the views and opinions expressed herein should not be construed to be those of the National Institutes of Health.

The authors are thankful to the anonymous referees for valuable suggestions.

REFERENCES

[1] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "A framework for efficient data anonymization under privacy and accuracy constraints", *ACM Transactions on Database Systems*, 2009, Vol. 34, No. 2, Article 9.

[2] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[3] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, 2012.

[4] H. T. Nguyen and E. A. Walker, *First Course In Fuzzy Logic*, CRC Press, Boca Raton, Florida, 2006.

[5] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2007.

[6] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-Based System*, 2002, Vol. 10, No. 5, pp. 557–570.

[7] G. Xiang and V. Kreinovich, "Data anonymization that leads to the most accurate estimates of statistical characteristics", *Proceedings of the IEEE Series of Symposia on Computational Intelligence SSCI'2013*, Singapore, April 16–19, 2013, to appear.

[8] L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.