Quantifying the contribution of individual records to the reidentification risk of (pseudo)anonymized datasets

Michel Béra[1], Vasiliki Daskalaki[2], Gilbert Saporta[3], Kimon Spiliopoulos[2], Konstantinos Spinakis[4] and Photis Stavropoulos[2]

[1] Professor emeritus of CNAM, Chair of Statistical Modelling of Risk; Founding member, Team ESD R3C; Associate researcher, Lab CEDRIC/MSDMA
michel.bera@lecnam.net

[2] Quantos SA Statistics and Information Systems, Syggrou 154, 17671 Athens, Greece
{vasiliki.daskalaki, k.spiliopoulos, photis.stavropoulos}@quantos-stat.com

[3] Professor emeritus of CNAM, Chair of Applied Statistics; Researcher, Lab CEDRIC/MSDMA
{gilbert.saporta@cnam.fr}

[4] MSc student in Applied Mathematics, Ecole Polytechnique Fédérale de Lausanne
konstantinos.spinakis@epfl.ch

Abstract. The reidentification of individuals or business establishments in (pseudo)anonymized microdata may expose sensitive data and will lead to fines and reputational damage for the data's custodians. The QaR method (AFNOR, 2020) proposes a measure of the reidentification risk of a dataset, and a statistical technique, based on extreme-value theory, to estimate it. This risk has great value. It is a gauge of the effectiveness of whatever disclosure control the custodians apply to the data; it could be reported to regulatory authorities to demonstrate the custodians' level of care for the data subjects' privacy; it can be used to calculate an insurance premium against unauthorized disclosure or the amount of money that custodians need in their balance sheet to cover potential financial damages due to such disclosure.

The present paper deals with a particular aspect of the methodology: the quantification of the contribution of each record to the dataset's risk. It discusses its importance and its large computational demands in very large datasets, and proposes metrics that are faster to compute and could serve as proxies of record contribution. The results for some of these proxies are promises but more investigation is needed.

Keywords: Anonymization, extreme-value theory, privacy, pseudonymization, reidentification, risk.

1 Introduction

Personal data about large segments of the population (be it humans or legal entities) proliferate in public and private authorities. They are recognized as a valuable

commodity for policy and business purposes. Unfortunately, they are also valuable for malevolent purposes and are frequently targets of attempts to steal them. Legislation is very demanding: the custodians of personal / business data databases can face severe penalties, as well as reputational damage, if data are stolen. Consequences are more severe if the identities of individual persons / entities are revealed. Custodians are obliged to protect the identity of the subjects of the data, irrespectively of whether the data are intended to be shared or not with authorities or researchers.

Anonymization and *pseudonymization* are two classes of measures for protecting data subjects' identities. While anonymization is considered irreversible and pseudonymization reversible, the proliferation of huge amounts of data generated and collected by devices and software tools pushes the boundary between anonymized and pseudonymized data towards the latter. What is anonymous today may not be anonymous tomorrow.

According to this view, which we share, all datasets carry reidentification risk. The QaR method (AFNOR, 2020) proposes a measure of the reidentification risk of a dataset and a statistical technique, based on extreme-value theory, to estimate it. This risk has great value. It is a gauge of the effectiveness of whatever disclosure control the custodians apply to the data; it could be reported to regulatory authorities to demonstrate the custodians' level of care for the data subjects' privacy; it can be used to calculate an insurance premium against unauthorized disclosure or the amount of money that custodians need in their balance sheet to cover potential financial damages due to such disclosure.

Béra et al. (2022) present software they have developed to implement the QaR method and identify directions for future research related to the method. Investigation along one of these directions, the quantification of the contribution of each record to the dataset's risk are discussed in this paper.

Section 2 presents the QaR method. Section 3 describes the exact computation of record's contribution to dataset risk, its significance, and its computational demands. It also presents four proxies that have been investigated and their results on a test dataset. The results are promising but additional investigation is needed; these needs are outlined in section 4.

2 The QaR method

Consider a (pseudo)anonymized dataset consisting of L records and N variables. Each record contains data on a single individual (person or company) and each column corresponds to one attribute (e.g., sex, age, last year's total income, country of birth etc.).

To identify individuals, an intruder will try to match records of the dataset with records in not anonymous datasets to which the intruder has access. Attempts to matching will be made by examination of combinations of variables in the (pseudo)anonymized dataset, identification of records with unique values for these combinations and identification of the exact same combinations of values in the not anonymous datasets.

The existence of records with unique combinations of values in the (pseudo)anonymized dataset is not a sufficient condition for reidentification. It is, however, a necessary one. The more such records are in the (pseudo)anonymized dataset the greater its reidentification risk is.

We call each set of p columns, with indices $j_1, j_2, \dots, j_p$, where $1 \leq j_1 < j_2 < \cdots, < j_p \leq N$, a *quasi-identifier of size p*. We denote this quasi-identifier as $Q(j_1, j_2, \dots, j_p)$, or, for brevity, $Q(j), j = 1,2, \dots, N_p$, where $N_p = \binom{N}{p}$. We associate with each quasi-identifier its *reidentification risk*

$$\theta(j) = \theta\big[Q(j_1, j_2, \dots, j_p)\big] = \frac{H\big[Q(j_1, j_2, \dots, j_p)\big]}{L} = \frac{H(j)}{L}$$

where $H(j)$ is the number of distinct values assumed by the quasi-identifier in the dataset.

Clearly, $0 \leq \theta \leq 1$. The closer θ is to 1, the more unique records exist in the dataset with respect to their values for the quasi-identifier. Therefore, the greater the risk that individuals will be reidentified based on this quasi-identifier.

Taking in turn each of the quasi-identifiers of size p formed by the columns of the dataset and computing the reidentification risk associated with each one, we obtain a set of risks $\theta(j), j = 1,2, \dots, N_p$.

The QaR method treats this set as a random sample from a distribution of risks and defines the reidentification risk of the dataset as an upper quantile of this distribution.

More precisely, *the reidentification risk, $T(p, \alpha)$, of the dataset is the $1 - \alpha$ quantile of the distribution of risks associated with the quasi-identifiers of size p*. The interpretation of the measure is as follows: the probability that a quasi-identifier of size p will have a reidentification risk associated with it that is larger than $T(p, \alpha)$ is α.

The estimation of $T(p, \alpha)$ relies on extreme value theory. Its estimate is not an empirical quantile of the computed θ; it is, instead, an estimate obtained by fitting a distribution on a monotonic transformation of the most extreme θ.

More precisely, the empirical $1 - \pi_u$ quantile, u, of the θ is computed. This empirical quantile is computed with definition 8 of Hyndman and Fan (1996).

Only those θ which satisfy $\theta > u$ are retained and a monotonic transformation of them is computed, as follows:

$$l(j) = \begin{cases} \ln\left(\frac{\theta(j)}{1-\theta(j)}\right) - \ln\left(\frac{u}{1-u}\right) & , \quad \text{if } \theta(j) > u \\ \text{NULL} & , \quad \text{if } \theta(j) \leq u \end{cases}$$

A Generalised Pareto Distribution (GPD henceforth; see McNeil et al (2005), sec. 7.2) is fitted on the $l(j)$ which are not NULL. The reader is reminded that the GPD has cumulative distribution function (c.d.f.)

$$P(X \leq x) = G_{\xi,\beta}(x) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\beta}\right)^{-\frac{1}{\xi}}, \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\beta}\right), \xi = 0 \end{cases}$$

where $\beta > 0$, $x \geq 0$ when $\xi \geq 0$, and $0 \leq x \leq -\beta/\xi$ when $\xi < 0$.

Estimates of scale β and shape ξ are obtained with the method of probability-weighted moments; see Hosking and Wallis (1987). The estimator of $T(p,\alpha)$ is given by

$$\hat{T}(p,\alpha) = exp\left(ln\left(\frac{u}{1-u}\right) + \frac{\hat{\beta}}{\hat{\xi}}\left[\left(\frac{\alpha}{\pi_u}\right)^{-\hat{\xi}} - 1\right]\right) \Big/ \left\{1 + exp\left(ln\left(\frac{u}{1-u}\right) + \frac{\hat{\beta}}{\hat{\xi}}\left[\left(\frac{\alpha}{\pi_u}\right)^{-\hat{\xi}} - 1\right]\right)\right\}$$

Parameters p, α and π_u are a choice of the user. Usual values are $p = 3$, $\alpha = 0.01$ or $\alpha = 0.001$, and $\pi_u = 0.05$.

3 The contribution of a record to dataset risk and proxies of it

The contribution of a record to the dataset's risk is the change in the risk due to the removal of the record from the dataset with all other records remaining the same. We denote the risk of the dataset without record i, $i = 1,2,\dots,L$, as $\hat{T}(p,\alpha{:}\,i)$ and the contribution of record i to the risk as $DT(i) = \hat{T}(p,\alpha) - \hat{T}(p,\alpha{:}\,i)$.

The quantification of records' contribution serves many purposes. The withdrawal of a few with the largest contributors may render the dataset a lot safer. Furthermore, the contribution of a record can be considered as a proxy of the risk of identifying the statistical unit this record corresponds to. Finally, the investigation offers additional insights into the properties of the QaR method itself.

A 'proper' backward elimination of records would require that at each step, say s,

- the record with the highest contribution, say i_s, is removed from the dataset,
- the dataset's risk becomes $\hat{T}_s(p,\alpha) = \hat{T}_{s-1}(p,\alpha{:}\,i_s)$, where $\hat{T}_0(p,\alpha) = \hat{T}(p,\alpha)$, the original dataset's risk, and
- the process moves to the next step.

This process can be very expensive computationally in very large datasets. Suppose that one wants to remove S records. The process requires the computation of dataset risk for

$$1 + \sum_{s=1}^{S}(L - s + 1)$$

different datasets. Once for the original dataset, and as many times as the records still in the dataset at each subsequent step. It may even be very expensive to compute $DT(i) = \hat{T}(p,\alpha) - \hat{T}(p,\alpha{:}\,i)$ only for the first step of the process.

Proxies to DT can substitute for the computation of DT in a backward elimination process. They can even be computed only on the original dataset and by just those values indicate which block of records to remove without making distinct elimination steps.

3.1 Test data and experimental setup

A small, amenable dataset was used in the tests presented in this paper. It is an extract from edition 1.3 of Round 10 of the European Social Survey (ESS10 henceforth), consisting of variables that are of ordinal scale at least. Records containing special values were removed. The final dataset consists of $L = 12129$ records and $N = 18$ variables. Its risk was computed as $\hat{T}(3,0.01) = 0.1948661$.

The contribution of each record to this risk was computed and will be examined together with the different discussed proxies.

3.2 Contribution to the original dataset risk as an indicator of the records to remove in backward elimination

To avoid backward elimination's demanding computational requirements, one can take the computed contributions $DT(i)$ of the records to the original dataset's risk, sort the records in decreasing order of contribution and remove those with the largest ones.

This implies the assumption that the record with the sth highest contribution to the original dataset's risk is the record that would be selected at step s of a proper backward elimination.

The following figure shows the impact on the dataset's risk of removing records in blocks of 100, from removing the top 100 most contributing ones up to removing the top 2500 most contributing ones, i.e., approximately 20% of the records. We consider that removing more than 20% of the records would render a dataset useless for the purposes for which it was created.

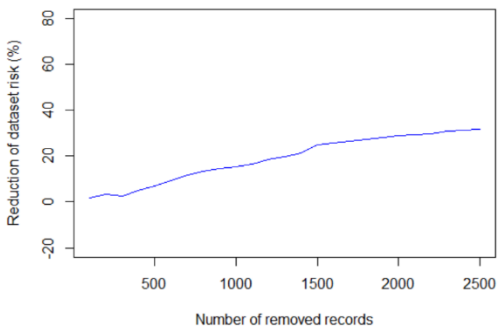

Fig. 1. Reduction of the ESS10 dataset risk, as a proportion of the original dataset's risk, due to the removal of records based on their contribution $DT(i)$ to the original risk.

It can be observed that removing the 2500 most contributing records to the original dataset's risk leads to a dataset whose risk is almost 32% smaller than the original ones. 'Along the way', however, the reduction does not follow a monotonic pattern. For

instance, removing the top 200 records leads to a reduction of 3.5%, while removing the 300 top records leads to a reduction of 2.7%. In other words, the dataset with the 300 top records removed has a larger risk than the dataset with 200 records removed.

3.3 A record's uniqueness pattern as proxy for its contribution to risk

Let

$$U(i,j) = \begin{cases} 1 & \text{if record } i \text{ is unique with respect to quasi-identifier } j \\ 0 & \text{if record } i \text{ is not unique with respect to quasi-identifier } j \end{cases}$$

$\theta(j:i)$ be the reidentification risk of quasi-identifier j when record i has been removed from the dataset, and
$H(j:i)$ be the number of distinct values assumed by the quasi-identifier in the dataset when record i has been removed from it. It can be expressed as a function of $U(i,j)$ and $H(j)$:

$$H(j:i) = U(i,j)[H(j) - 1] + [1 - U(i,j)]H(j).$$

The contribution of the record to the risk of the quasi-identifier is computed as follows:

$$\begin{aligned} \theta(j) - \theta(j:i) &= \frac{H(j)}{L} - \frac{H(j:i)}{L-1} \\ &= \frac{LH(j) - H(j) - L\{U(i,j)[H(j) - 1] + [1 - U(i,j)]H(j)\}}{L(L-1)} \\ &= \frac{LU(i,j) - H(j)}{L(L-1)} \end{aligned}$$

However, $H(j) = L\theta(j)$ and therefore,

$$\theta(j) - \theta(j:i) = \frac{U(i,j) - \theta(j)}{L-1} \Leftrightarrow \theta(j:i) = \theta(j) + \frac{\theta(j) - U(i,j)}{L-1}$$

The following observations can be made:

- If the record is unique with respect to the quasi-identifier, its removal decreases the quasi-identifier's risk.
  - The smaller the quasi-identifier's risk in the complete dataset, the greater will be the risk reduction by removal of the unique record.
  - If a record is unique with respect to all quasi-identifiers, its removal will move all $\theta(j)$, and therefore the threshold u too, downwards. It will, arguably, lead to a reduction of the dataset's risk.
- If the record is not unique with respect to the quasi-identifier, its removal increases the quasi-identifier's risk.
  - The greater the quasi-identifier's risk in the complete dataset, the greater will be the risk increase by removal of the not unique record.

— If a record is not unique with respect to any quasi-identifier, its removal will move all $\theta(j)$, and therefore the threshold u too, upwards. It will, arguably, lead to an increase of the dataset's risk.

How do we choose between two records which are unique with respect to the same number of quasi-identifiers but not for the same quasi-identifiers? We examine this by using the two following record metrics:

$$U^*(i) = \sum_j U(i,j)[1 - \theta(j)]$$

and

$$U^+(i) = \sum_j U(i,j)\theta(j)$$

U^* favors records which are unique for quasi-identifiers with small risk, while U^+ favors records which are unique for quasi-identifiers with large risk.

Computing U^* and U^+ on the ESS10 dataset shows strong positive correlation between them. The Pearson correlation coefficient is 0.9490, while the Kendall one is 0.9248.

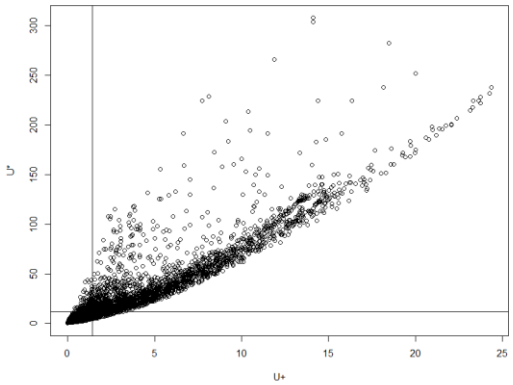

Fig. 2. Values of U^* and U^+ for the records of the ESS10 dataset. The lines represent the average values of U^* and U^+.

A second remark is that 6273 records, 51.7% of the total, have $U^*(i) = U^+(i) = 0$. More than half of the records are not unique for any quasi-identifier of size 3, which is consistent with the dataset's relatively low $\hat{T}(3,0.01)$.

Given the strong correlation, we examine the relationship of $DT(i)$ with $U^+(i)$ only.

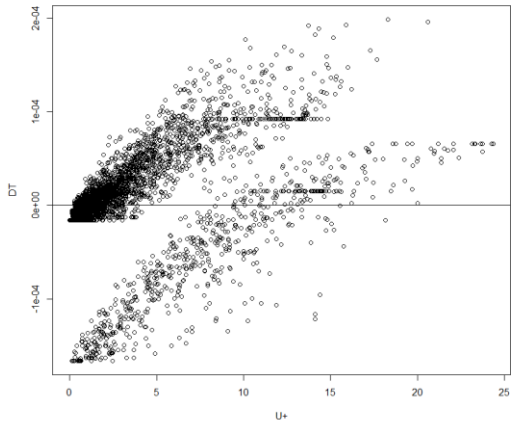

Fig. 3. Values of U^+ and DT for the ESS10 dataset.

The correlation between U^+ and the contribution of a record to the risk of the dataset is strong. The Pearson correlation coefficient is 0.5205, while the Kendall one is 0.6628.

What is remarkable in the plot is the appearance of two clusters of points, very similar in shape and well separated from each other. The correlation between DT and U^+ in either cloud would arguably be even higher than in the dataset as a whole. At the moment of writing these lines the investigation about the causes of the clustering or whether it appears in other datasets is ongoing.

The results indicate that U^+ can be a good indicator of records that will reduce the dataset risk when removed. This is corroborated by the two figures that follow.

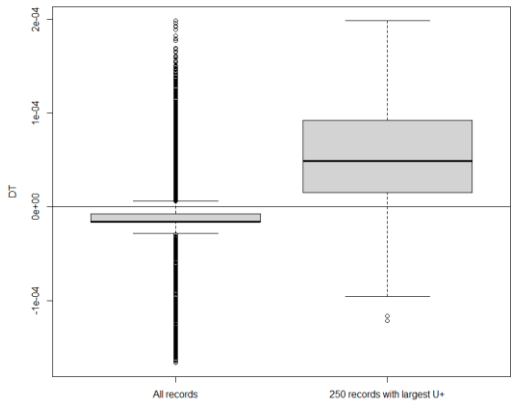

Fig. 4. Boxplots of DT in the ESS10 dataset.

Most of the 250 records with the highest values of U^+ (**Fig. 4**) would lead to a reduction of risk if removed, and a much greater one than the majority of records.

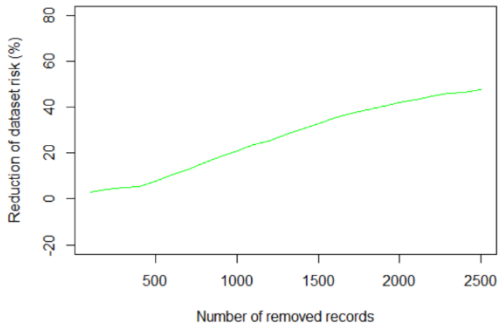

Fig. 5. Reduction of the ESS10 dataset risk, as a proportion of the original dataset's risk, due to the removal of records based on their value of U^+.

It can be observed (**Fig. 5**) that removing the 2500 records with the largest values of U^+ leads to a dataset whose risk is almost 48% smaller than the original one's. Moreover, each additional block of 100 records removed leads to additional reduction of the dataset's risk.

3.4 Record entropy as proxy for its contribution to risk

Consider a record $i, i = 1,2, \dots, L$ and a quasi-identifier $j, j = 1,2, \dots, N_p$. Let the record's values for this quasi-identifier be $x_{i,j_1}, x_{i,j_2}, \dots, x_{i,j_p}$.

Compute the value of the empirical p-variate cumulative distribution function of the quasi-identifier for this record:

$$\mathrm{F}(i,j) = \frac{1}{L}\sum_{r=1}^{L} \mathbb{1}\left(x_{r,j_1} \le x_{i,j_1}, \quad x_{r,j_2} \le x_{i,j_2}, \dots, \quad x_{r,j_p} \le x_{i,j_p}\right)$$

where $\mathbb{1}$ is the indicator function.

As an extra step of standardization, take the empirical cumulative distribution function of the computed p-variate function:

$$\mathrm{H}(i,j) = \frac{1}{L}\sum_{r=1}^{L} \mathbb{1}\left(\mathrm{F}(r,j) \le \mathrm{F}(i,j)\right)$$

Subsequently, split interval $[0,1]$ into $\sqrt{N_p}$ intervals $I_m, m = 1, 2, \dots, \sqrt{N_p}$, where $I_m = [(m-1) * 1/\sqrt{N_p}\ ,\ m * 1/\sqrt{N_p}\ ]$. The intervals have width $1/\sqrt{N_p}$.

For each record i define p_m^i as the proportion of values of $\mathrm{H}(i,j), j = 1,2, \dots, N_p$, that fall into I_m.

We then compute the entropy $E(i)$ of each record i as

$$E(i) = \sum_{m=1}^{\sqrt{N_p}} E_m^i$$

with

$$E_m^i = \begin{cases} -\, p_m^i \, log(p_m^i), & p_m^i > 0 \\ 0 \quad , & p_m^i = 0 \end{cases}$$

The following diagram shows the relationship between $DT(i)$ and $E(i)$ on the ESS10 dataset.

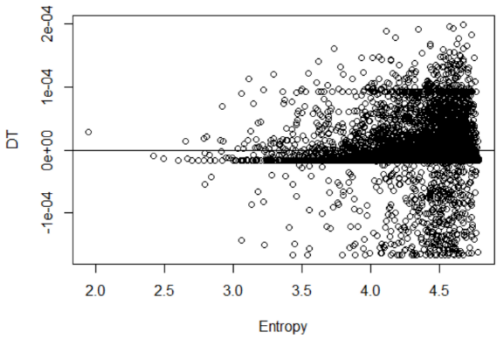

Fig. 6. Values of E and DT for the ESS10 dataset.

A trumpet like shape is evident. The entropy of the records has no evident relationship with their contribution to the dataset's risk. The lack of relationship between contribution to risk and entropy is explained by the fact that a record that is unique for several quasi-identifiers may have very small values in all of them (small entropy), very large values in all of them (small entropy, again) or large – average – small values depending on the quasi-identifier (large entropy).

The following figure even shows a negative relationship with risk.

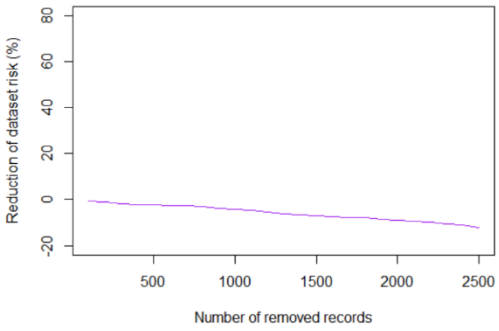

Fig. 7. Reduction of the ESS10 dataset risk, as a proportion of the original dataset's risk, due to the removal of records based on their entropy.

It can be observed that removing records based on their entropy leads to negative reduction, i.e., to increase of the dataset's risk. Removing the 2500 records with the largest entropies leads to a dataset with 12% largest risk.

An additional disadvantage of entropy is that it requires that the variables reported in the dataset are at least ordinal. With nominal variables no empirical cumulative distribution can be computed.

3.5 Record contents as predictors of its contribution to risk

In this section we examine whether the values of the variables in a record can predict the record's contribution to dataset risk.

Partial least squares (PLS) regression with the records' DT as response variable and the 18 variables of the dataset as explanatory variables gave a model with a very poor R^2.

For this reason, PLS regression was then applied with a binary response variable:

$$S(i) = \begin{cases} 1, & DT(i) > 0 \\ 0, & DT(i) < 0 \end{cases}$$

This binning turns the approach into a classification problem.

The number of factors for the PLS regression is 3. We tested the robustness of the results by selecting at random 75% of the records for training and keeping the remaining 25% records for testing.

The AUC for the two models (test vs training) remains stable: 0.7063 on test data, 0.7059 on training data.

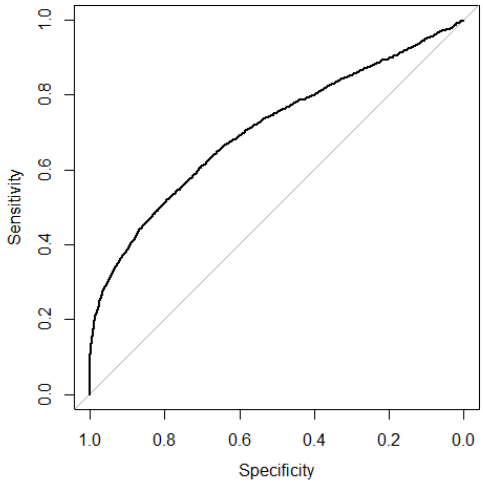

Fig. 8. Performance of PLS regression as a tool for predicting the sign of a record's contribution to dataset risk.

The following figure shows the impact on dataset risk of the removal of records in blocks of 100, based on their contribution to dataset risk as predicted by PLS regression.

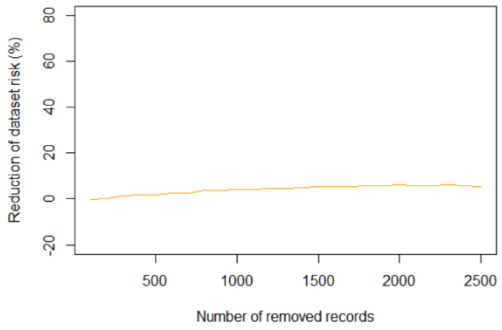

Fig. 9. Reduction of the ESS10 dataset risk, as a proportion of the original dataset's risk, due to the removal of records based on their contribution to dataset risk as predicted by PLS regression.

The result shows a positive but limited effect. Removal of the top 2500 records leads to a dataset whose risk is 5% smaller than the original one's.

4 Conclusions

The records' contribution to the original dataset risk or the records' U^+ metric on the original dataset give quite good indication of which records would be removed in a proper backward elimination. They provide large computational gains as the former requires only $L + 1$ computations of dataset risk, while the latter requires none.
The entropy and the PLS regression-based metrics, on the other hand, do not help identify impactful records to remove. The following figure brings together the performance of these metrics.

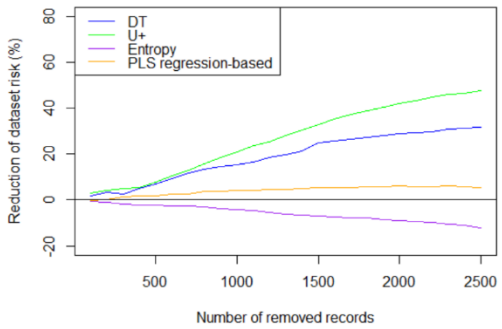

Fig. 10. Reduction of the ESS10 dataset risk, as a proportion of the original dataset's risk, due to the removal of records, selected according to several alternative criteria.

These results remain to be verified in additional and larger datasets. Several additional investigations remain to be made:

- Select blocks of records to remove with proper backward elimination and with computation of U^+ in each step and compare the change of dataset risk with that achieved when selecting the blocks to remove according to the initial values of U^+.
- Study theoretically the impact of a record's removal on quasi-identifier risks and dataset risk. For instance, can one set an upper bound on the reduction of risk that can be achieved by removing a record or by removing S records? The computation of such a bound would help assess whether it is worthwhile to attempt removing records.
- Understand the emergence of the clusters in the plot of DT versus U^+.
- Find ways to recompute quickly matrix U after each record's removal. This matrix is a component of the computation of U^+ and of the the quasi-identifiers' new risks. Its quick computation could speed up the execution of proper backward elimination.
- Attempt to improve the performance of the PLS regression-based metric by fitting a quadratic polynomial of the dataset variables. In relation to this, try to combine entropy and PLS regression by defining a 'filtered entropy' being, e.g., equal to 0 for records predicted by PLS regression to increase risk and equal with the original entropy for records predicted by PLS regression to reduce risk.
- Search for other record metrics which are more highly correlated with risk reduction.

References

1. AFNOR (2020) AFNOR SPEC Z90-030: La méthode QaR de mesure du risque extrême de réidentification d'une base de données dans le cadre de l'évaluation de son assurabilité. https://bivi.afnor.org/notice-details/afnor-spec-z90-030-fevrier-2020-risques-et-bases-de-donnees/1314498
2. Béra, M., Spiliopoulos, K., Spinakis, A. and Stavropoulos, P. (2022) Measuring reidentification risk in (pseudo)anonymized datasets: the QaR method and software. *Privacy in Statistical Databases (PSD) 2022*, Paris, September 21-23, USB-key Conference Proceedings. Available by the authors upon request.
3. Hosking, J. R. M., Wallis, J. R.: Parameter and Quantile Estimation for the Generalized Pareto Distribution. Technometrics 29(3), 339-349 (1987).
4. Hyndman, R. J., Fan, Y.: Sample Quantiles in Statistical Packages, Statistical Computing 50(4), 361-385 (1996).
5. McNeil, A. J., Rüdiger, F., Embrechts, P.: Quantitative Risk Management, Concepts, Techniques and Tools. Princeton University Press, Princeton New Jersey (2005).