UNECE 2013

On the Hellinger distance for measuring information loss in microdata

Vicenç Torra[1] and Michael Carlson[2]

October, 2013

[1] Institut d'Investigació en Intel·ligència Artificial (IIIA-CSIC), Bellaterra (Catalonia, Spain)

[2] Department of Statistics, Stockholm University, Stockholm (Sweden)

Introduction

Discussion on some measures: for data protected sets

- Hellinger distance
- Entropy
- Variation of entropy

Comparison: with other information loss and disclosure risk measures

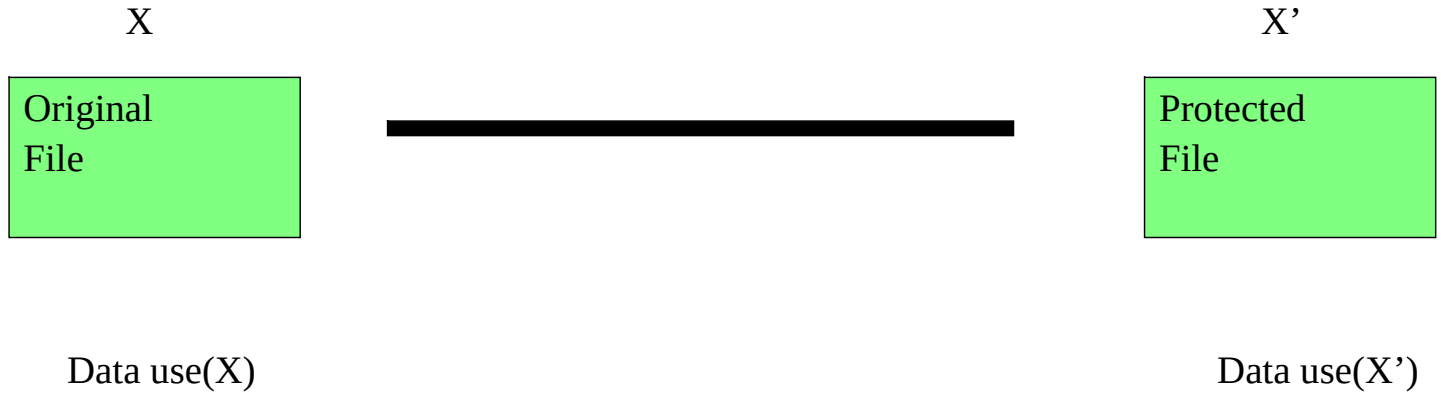
Outline

Setting

Data release: Given X release X'

Information loss:

- Similar results on the original file and protected file
- Information loss as a measure of the divergence between the results



- $IL = divergence(dataUse(X), dataUse(X'))$

Disclosure risk:

- Record linkage

Setting

Information loss:

- $IL = divergence(dataUse(X), dataUse(X'))$

Setting

Information loss:

- $IL = divergence(dataUse(X), dataUse(X'))$
- Data use: Probability distributions

Setting

Information loss:

- $IL = divergence(dataUse(X), dataUse(X'))$
- Data use: Probability distributions
- Divergences: Comparison of probability distributions
  - Hellinger distance (a f-divergence)
  - Entropy (also used as a *disclosure risk measure*)

Definitions

Definitions:

- Hellinger distance. Discrete probability distributions $P = (p_1, \ldots, p_K)$ and $Q = (q_1, \ldots, q_K)$:

$$H(P, Q) = \sqrt{\frac{1}{2} \sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2}.$$

Definitions

Definitions:

- (DR-)Entropy. Frequency counts $F = (F_1, \ldots, F_K)$ with $N = \sum_{i=1}^{K} F_i$:

$$H(\frac{F}{N}) = \frac{N \log N - \sum_{i=1}^{K} F_i \log F_i}{N}.$$

- (DR-)Entropy (Shlomo, Antal and Elliot). $w_1, w_2 \leq 0$, $w_1 + w_2 \leq 1$:

$$\begin{aligned} R(F, w_1, w_2) &= w_1 \left[\frac{\sum_{i=1}^{K} |F_i - \frac{1}{2}| + \frac{1}{2}K - N}{K} \right] \\ &+ w_2 \left[1 - \frac{N \log N - \sum_{i=1}^{K} F_i \log F_i}{N \log K} \right] \\ &+ (1 - w_1 - w_2) \left[\frac{1}{\sqrt{N}} \log \frac{1}{e\sqrt{N}} \right]. \end{aligned}$$

Experiments

Experiments: Same files in (Domingo-Ferrer, Torra, 2001)

- Original file: 1080 records, 13 numerical variables

Experiments

Experiments: Same files in (Domingo-Ferrer, Torra, 2001)

- Original file: 1080 records, 13 numerical variables
- Data protection: Distr, JPeg, Sampling, Additive noise, Rank Swapping, Microaggregation
- Different parameters

Experiments

Experiments: Same files in (Domingo-Ferrer, Torra, 2001)

- Original file: 1080 records, 13 numerical variables
- Data protection: Distr, JPeg, Sampling, Additive noise, Rank Swapping, Microaggregation
- Different parameters
- Total: 215 protected files

Experiments

Experiments: Same files in (Domingo-Ferrer, Torra, 2001)

- Original file: 1080 records, 13 numerical variables
- Data protection: Distr, JPeg, Sampling, Additive noise, Rank Swapping, Microaggregation
- Different parameters
- Total: 215 protected files

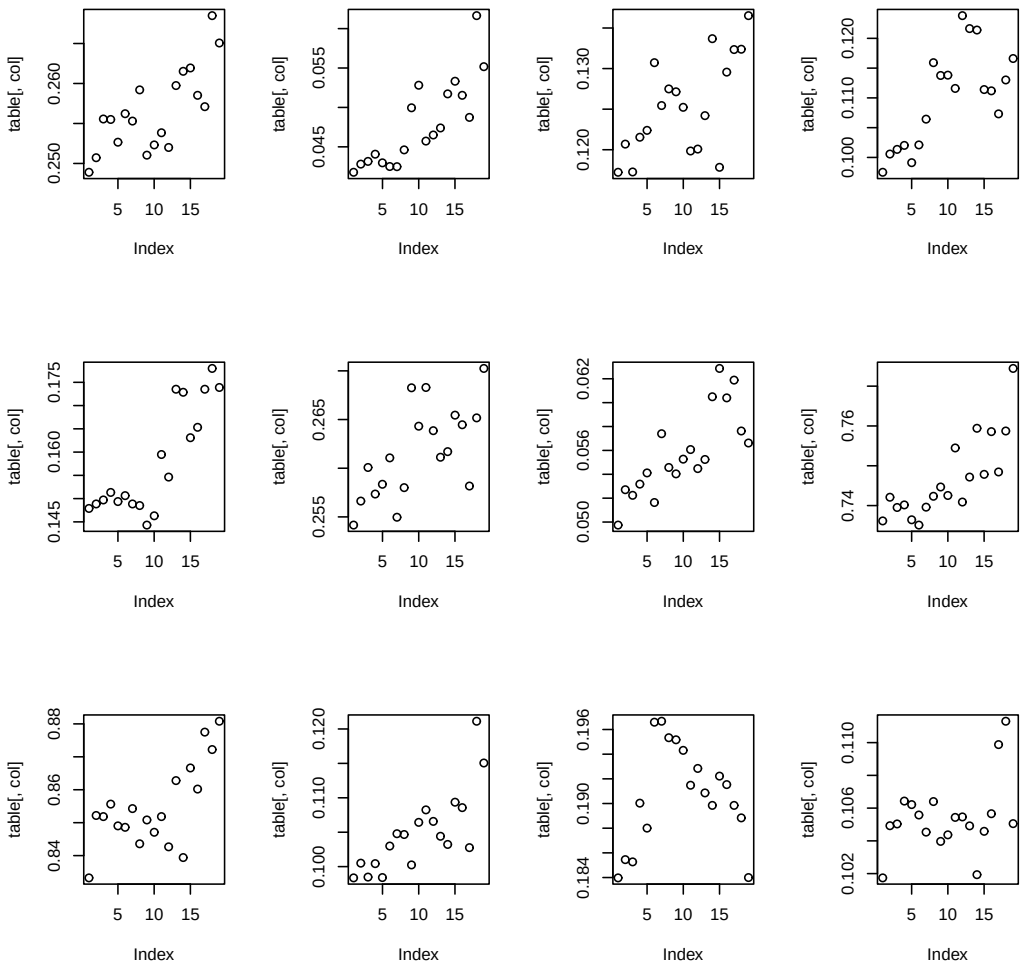
Measures:

- Measures above: Hellinger distance, Entropy, and Entropy-SAE
- Information Loss: PIL (statistics), FCM (clustering)
- Disclosure Risk: DBRL, MDBRL, PRL (record linkage)

Experiments

Results:

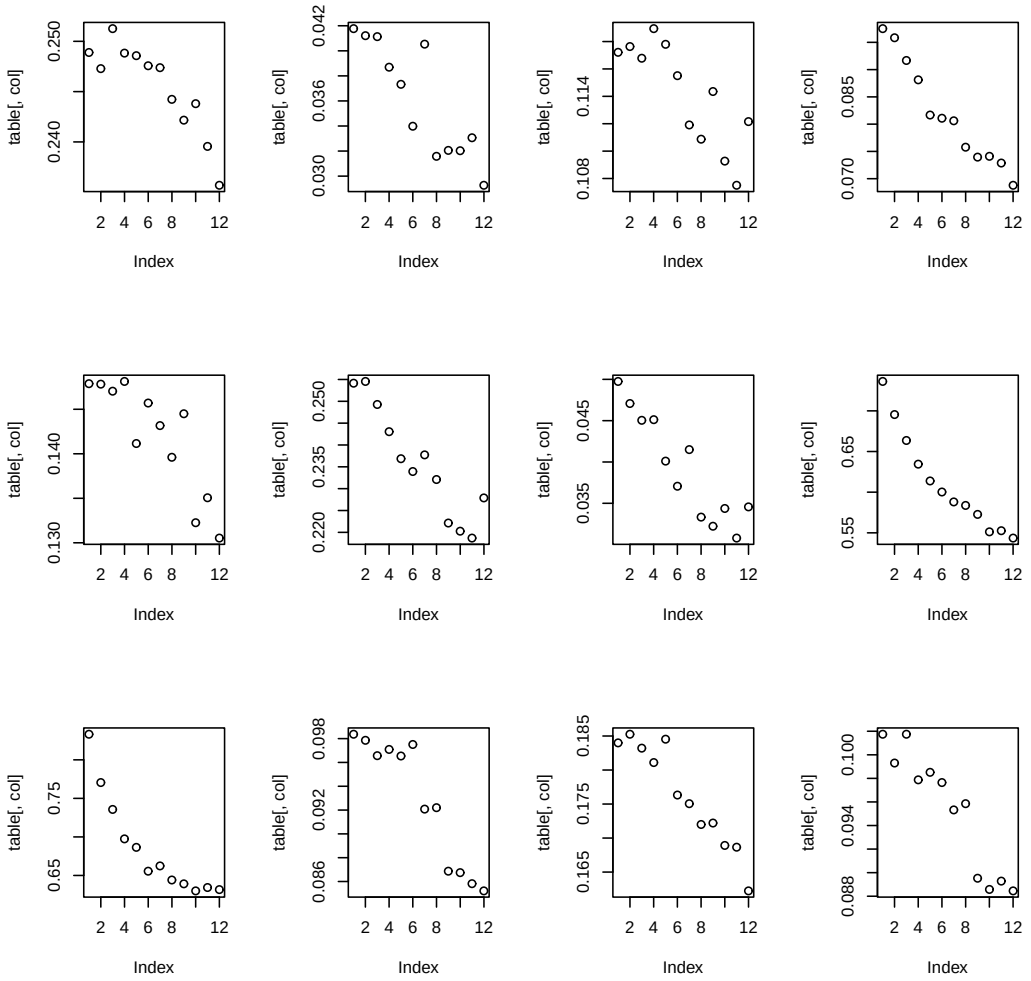
- Microaggregated file (entropy) (increasing)



Experiments

Results:

- Noise addition (entropy) (decreasing)



Experiments

Results:

- Correlation between Hellinger distance and the $entr$: 0.0576
- Correlation between Hellinger distance and the $entr.SAE$: 0.0104
- Correlation between the $entr$ and $entr.SAE$: 0.6341

Correlation coefficients:

	PIL	FCM	DBRL	MDBRL	PRL
hd	0.0569	-0.0109	-0.1603	-0.1616	-0.2153
$entr$	0.4778	0.6548	-0.4816	-0.4896	-0.3681
$entr.SAE$	0.3268	0.2869	-0.2495	-0.2507	-0.1906

Experiments

Entropy and data protection:

- Noise addition, entropy increases ($DR = 1 - entropy$, decreases)
- Microaggregation, entropy decreases ($DR = 1 - entropy$, increases)
- Rank swapping, entropy not modified (1D distribution unchanged)
- PRAM, entropy increases (with the typical markov matrix where diagonal is $0 < \Theta < 1$ and for rows probabilities add something different to zero)
- Global recoding, entropy decreases

Conclusion: When methods *concentrate* values entropy decreases, when methods *disperse* values entropy increases.

So: we need to define:

$$IL_1(original, protected) = |H(protected) - H(original)|.$$

Experiments

Experiments:

- Sample of 50 records of the original file
- Protection: microaggregation, rank swapping, noise addition (additive and correlated noise)
- Using sdcMicro (Templ, 2008)

Experiments

Results:

- Correlation:

	PIL	IL	DBRL	hd	2hd.v2	2hd.v3	entr	entr.SAE
PIL	1	0.4255	-0.585					
IL		1	-0.126					
DBRL			1					
hd	0.8549	0.4908	-0.079	1	0.953	0.986	0.854	0.365
2hd.v2	0.9271	0.4373	-0.317		1	0.967	0.825	0.377
2hd.v3	0.8472	0.4321	-0.109			1	0.868	0.428
entr	0.6314	0.5405	-0.044				1	0.721
entr.SAE	0.1264	0.3562	0.099					1

Summary

Summary and Conclusions

- Comparison of Hellinger and entropy with other IL and DR measures

Summary

Summary and Conclusions

- Comparison of Hellinger and entropy with other IL and DR measures
- 1D hellinger distance and entropies constant for rank swapping.
- Hellinger distance: increasing IL with increasing noise in microaggregation and noise addition (not perfectly monotonic)
- Entropy and Entropy-SAE: Increasing or decreasing according to the method.

Thank you