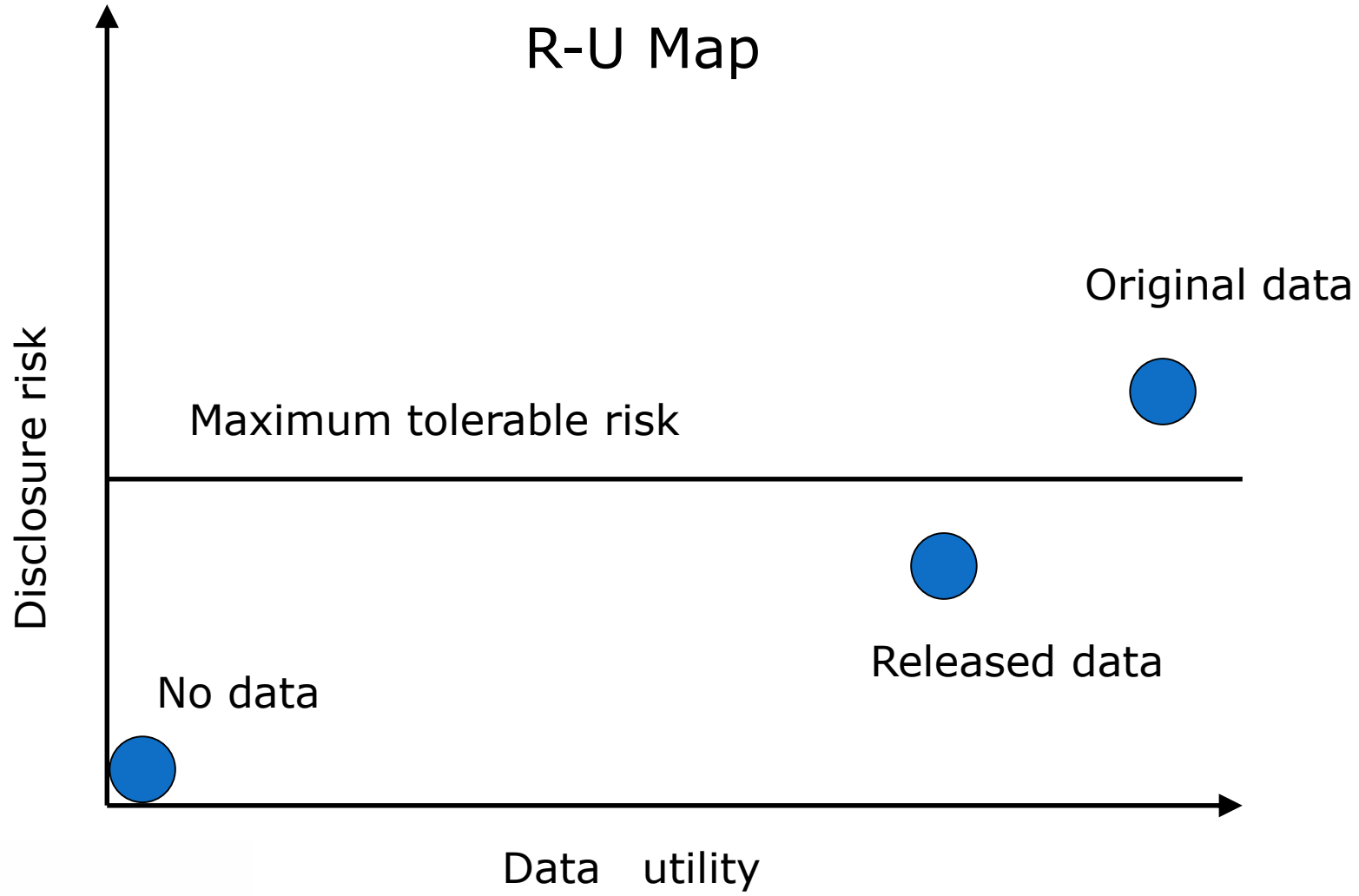
Utility evaluation of global recoding by accuracy of discrimination model

Statistical Research and Training Institute
Natsuki Sano

総務省統計研究研修所
Statistical Research
and Training Institute,
Ministry of Internal Affairs and Communications

Trade-off between Risk and Utility



総務省統計研究研修所
Statistical Research and Training Institute,
Ministry of Internal Affairs and Communications

Procedure of global recording (GR)

1. Aggregate category frequency for each variable and sort them.
2. Calculate thresh hold, $th = n \times p$, n:no.subject p:minimum frequency ratio.
3. If the category with minimum frequency is less than th, merge the category and second minimum category.

Category	D	B	F	…
Frequency	26	77	177	…

Merge

If there are plural categories with minimum frequency, merge them.

Category	D	A	C	…
Frequency	26	26	26	…

Merge

4. Repeat Step 3. until minimum frequency exceeds th

総務省統計研究研修所
Statistical Research and Training Institute,
Ministry of Internal Affairs and Communications

Information loss measure of categorical data

- Many survey item contain categorical variable e.g. census data and unsuitable to information loss measure of continuous variable
- Domingo-Ferror and Torra (2001) proposed the following measures for categorical data
  - Direct comparison of categorical value
  - Contingency table-based measure
  - Entropy-based measure

Evaluation of information loss by model accuracy (1)

- In many applied fields, some statistical model is applied to data including confidentiality protected data.
  - Used discrimination model: Multinomial logistic model
  - Model accuracy（ratio of correct prediction）are evaluated, before and after protecting by GR
- Hypotheses
  - When minimum frequency ratio p increases, information loss by model accuracy increases.
  - When the number of input variables increase, information loss by model accuracy increases.
  - High accuracy models in original data suffer from information loss by GR.

Evaluation of information loss by model accuracy (2)

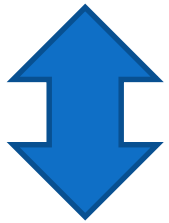
x_i:i-th valuable in original data　　x_i^G : i-th valuable after GR

$X_{sub}^{-i} \subseteq X^{-i} = \{x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_m\}$

$X_{sub,G}^{-i} \subseteq X_G^{-i} = \{x_1^G, x_2^G, \cdots, x_{i-1}^G, x_{i+1}^G, \cdots, x_m^G\}$

A: Original model ：prediction model by original data

$x_i = f(X_{sub}^{-i}),\ i = 1,2,\cdots,p$



Information loss: difference of accuracy between two models

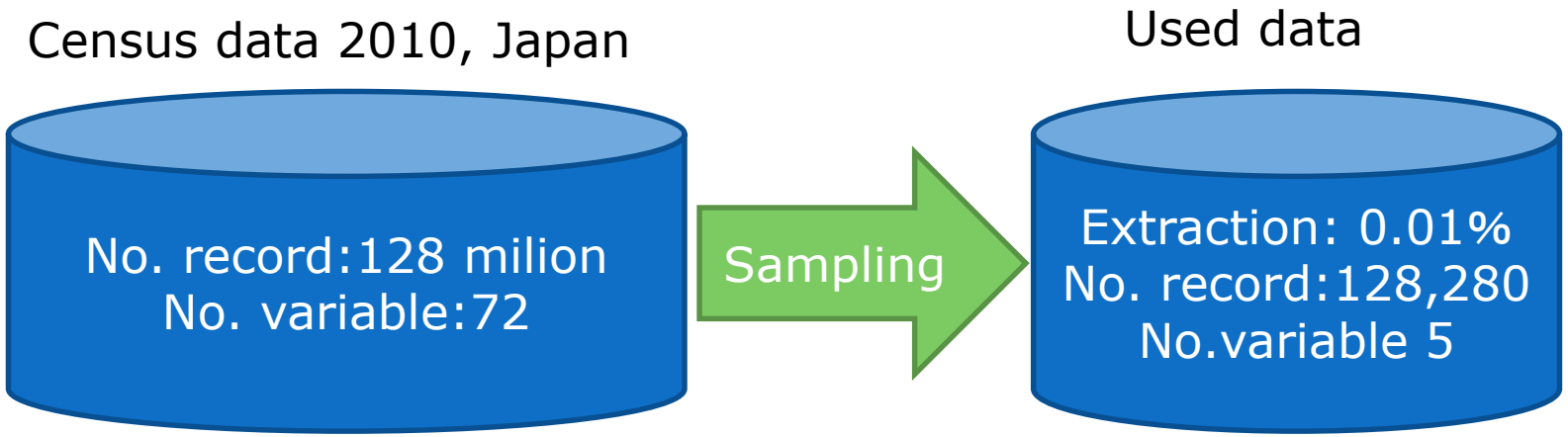
B:Global Recording model ： prediction model by GR data

$x_i = f(X_{sub,G}^{-i}),\ i = 1,2,\cdots,p$

Target varialble x_i in GR model is original variable.

総務省統計研究研修所
Statistical Research and Training Institute,
Ministry of Internal Affairs and Communications

Used data



Used 5 Variable

- Major division of industry
- Major division of occupation
- Type of family
- Classification of nationality
- Type by No. children and age

Preliminary analysis : Cramer's coefficient

	Major division of industry	Major division of occupation	Type of family	Classification of nationality	Type by No. children and age
Major division of industry	–	0.663	0.050	0.024	0.055
Major division of occupation	–	–	0.065	0.027	0.072
Type of family	–	–	–	0.023	0.271
Classification of nationality	–	–	–	–	0.020
Type by No. children and age	–	–	–	–	–

$$V = \sqrt{\frac{\chi_0^2}{n\{(\min(a,b)-1\}}}$$

n: no. record
a: no. category of row
b: no. category of column

総務省統計研究研修所
Statistical Research and Training Institute,
Ministry of Internal Affairs and Communications

Change of number of category by GR

p=Minimum frequency ratio in GR

Variables	Original	p=0.01	p=0.03	p=0.05
Major division of industry	22	17	10	7
Major division of occupation	13	12	9	8
Type of family	27	16	10	7
Classification of nationality	52	2	1	1
Type by No. children and age	97	29	12	8

Results of information loss (Mean value)

Mean	p=0.01	p=0.03	p=0.05
No. input variables = 1	0.000	0.007	0.010
No. input variables = 2	0.001	0.012	0.019
No. input variables = 3	0.001	0.018	0.029
No. input variables = 4	-0.000	0.024	0.041

- When minimum frequency ratio increases p, information loss increases (accuracy of discrimination model decreases).
- When the number of input variables increase, information loss increases (accuracy of discrimination model decreases).

Top 5 models with highest information loss (1)

No. input variables = 1, No. all models = 20, p=0.05

Model	A:Complete model	B:GR model	Information Loss	Cramer's coefficient
Occupation~Industry	0.812	0.727	0.085	0.663
No.Child & age~Family	0.540	0.481	0.059	0.271
Industry~Occupation	0.767	0.721	0.046	0.663
Family~No.Child & age	0.587	0.569	0.017	0.271
Family~Nationality	0.409	0.408	0.001	0.023

Top 10 models with highest information loss (2)

No. input variables = 2, No. all models = 30, p=0.05, Red:Large Cramer's V

Model	A:Complete model	B:GR model	Information Loss
Occupation~Industry+Nationality	0.812	0.727	0.085
Occupation~Industry+No.Child & age	0.809	0.731	0.078
Occupation~Industry+Family	0.813	0.736	0.076
No.Child & age~Family+Nationality	0.539	0.481	0.058
No.Child & age~Industry+Family	0.538	0.488	0.05
No.Child & age~Occupation+Family	0.537	0.487	0.049
Industry~Occupation+Nationality	0.767	0.721	0.045
Industry~Occupation+No.Child & age	0.76	0.72	0.039
Industry~Occupation+Family	0.76	0.73	0.029
Family~Industry+No.Child & age	0.593	0.571	0.023

Negative information loss (-0.006, Family~Industry)

Predicted class of Family (A:Original model) Accuracy=0.402

	01	02	03	04	05	06	07	08
0	3349	121542	0	0	0	0	0	0
10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
0	0	0	0	0	0	0	0	0
19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**VV**
0	0	0	0	0	0	0	3389	0

Predicted class of Family (B:GR model, p=0.05) Accuracy=0.408

	01	02	03	04	05	06	07	08
0	0	128280	0	0	0	0	0	0
10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
0	0	0	0	0	0	0	0	0
19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**VV**
0	0	0	0	0	0	0	0	0

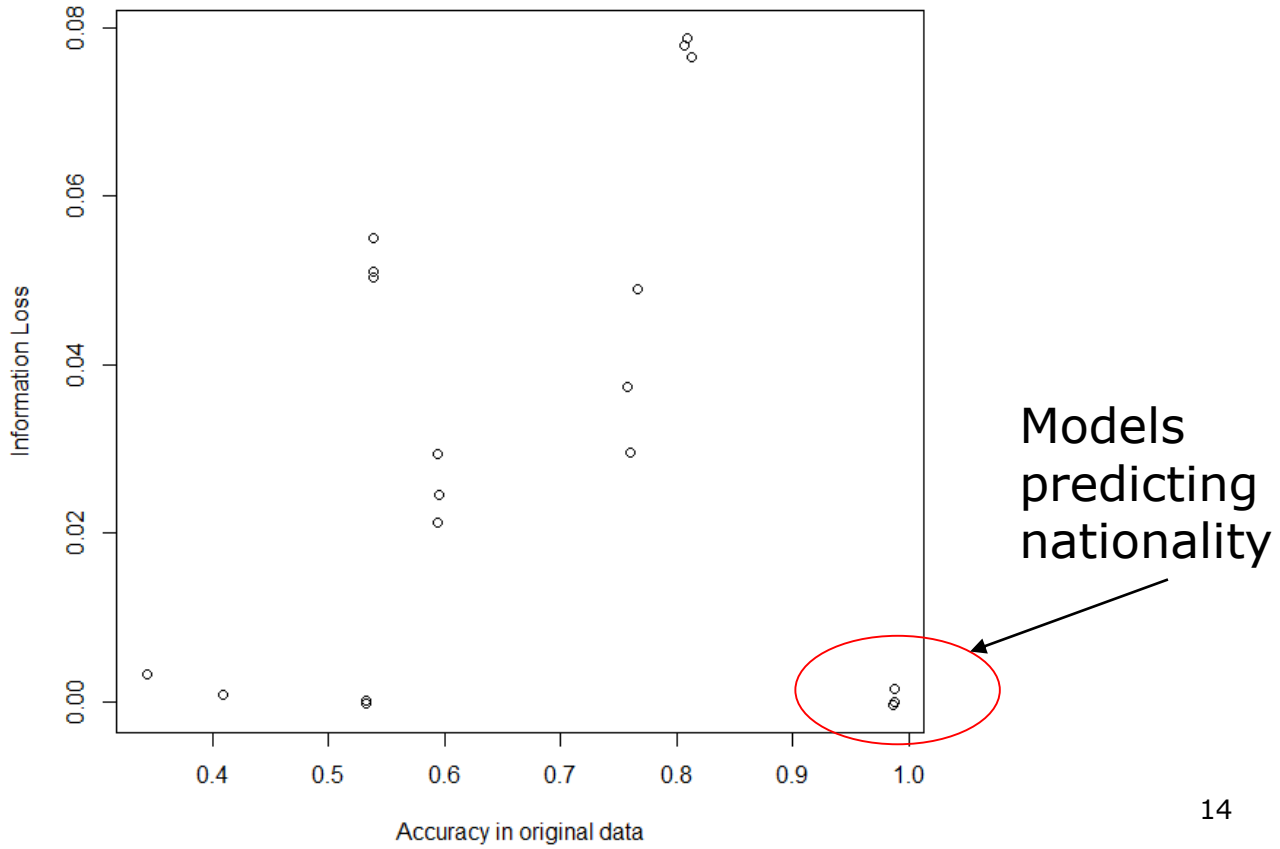
Accuracy in original data VS Information loss

No. input variables = 3, No. all models = 20, p=0.05

Apparently, as accuracy of original model increases, information loss increases

Correlation of all models : 0.017

Correlation except for models predicting nationality: 0.726



総務省統計研究研修所
Statistical Research and Training Institute,
Ministry of Internal Affairs and Communications

Conclusion

- We provide information to decide p in terms of loss of accuracy
  - In this case study, there is just about no information loss when p is 0.01, Even if p is 0.05 and the number of variables is 4, information loss is 0.041
- Knowledge from this simulation
  - When minimum frequency ratio increases p, information loss by model accuracy increases.
  - When the number of input variables increase, information loss by model accuracy increases ⇒ Large scale model needs to be dealt with cautiously.
  - High accuracy models in original data suffer from information loss by GR except for models predicting dominant category.

References

- Statistical disclosure control, Hundepool A., Domingo-Ferrer J., Franconi L., Giessing S., Nordholt E. S., Spicer K., et al. ,2012, John Wiley & Sons.
- Disclosure protection methods and information loss for microdata In confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies Domingo-Ferrer J. and Torra V., pp.91-110, North-Holland.

Thank you for your attention!