Using Generative Adversarial Network to Generate Synthetic Population

Yijun Wei[1]

Luca Sartore[1]

Nell Sedransk[1]

1 National Institute of Statistical Sciences. National Agricultural Statistics Service, United States Department of Agriculture

Disclaimer

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA, or U.S. Government determination or policy

Outline

- Background - Census of Agriculture
- Goal of the research
- Approach: Generative Adversarial Network (GAN)
- Experiment and result
- Conclusion and future work

Census of Agriculture

Every five years, USDA's National Agricultural Statistics Service (NASS) conducts the U.S. Census of Agriculture

- The Census provides a detailed picture of U.S. farms, ranches and the people who operate them
- It is the only source of uniform, comprehensive agricultural data for every state and county in the United States
- NASS makes Census data publicly available only as summary statistics
- Record-level information should be provided and disclosure of the confidential information should be averted

Goal of the research

- Problem: To provide detailed information based on the Census data
- Constraint: To avoid the disclosure of the confidential information
- Solution: A modified or synthetic dataset that preserves the internal relationship of the original dataset
- Previous approaches (Rubin, 1993; Reiter, 2005a,b,c; Paiva et al., 2014; Drechsler and Reiter, 2009):
  - Synthetic data distributions generated from models
  - Pooled or near-neighbors, used as exchangeable observations
  - Inter-changes of data elements among units
- The trade-off for synthetic data is disclosure protection vs preservation of data complexity

Motivation of Generative Adversarial Network (GAN)

Idea:

- To preserve finer internal structures
- To duplicate statistical properties with the original dataset

Approach:

- To use deep learning networks to synthesize
- To revise until synthetic data cannot be distinguished from the original data

Solution: Pair of networks

- Generative network (G-network) creates record-level synthetic data
- Discriminative network (D-network) distinguishes real data from the synthetic

Generative Adversarial Network (GAN)

GAN (Goodfellow et al., 2014) consists of two neural networks that train simultaneously and "compete" with each other

- G-network takes in random numbers and returns record-level synthetic data
- The generated synthetic data concatenated with the real data are fed into the D-network
- A D-network learns to distinguish real data from synthetic data
- Parameters of G-network are updated to "fool" the discriminator

Iterate

- Process stops when the G-network's output (synthetic data) can't be distinguished by D-network from the real population

Two utility measures

Two utility measures are used adapted from Woo et al. (2009)

Propensity score

- Assume: Original population and synthetic population have the same size
- Original records labeled 0, synthetic records labeled 1
- Propensity score for each record is generated by a model with response either 0 or 1

$$U_p = 1 - \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - c)^2$$

U_p: Propensity utility score

N: Total number of records

$\hat{p}_i$: Estimated propensity score for unit *i*

c: *0.5*

- Completely indistinguishable when $U_p \to 1$, otherwise $\to 0.75$

Two utility measures - continued

Clustering Score Measure

- Assume: Original population and synthetic population have the same size
- All the clusters are equally important

$$U_c = 1 - \frac{1}{G}\sum_{i=1}^{G} w_i\left(\frac{n_{io}}{n_i} - c\right)^2$$

U_c: Clustering utility score

G: Total number of clusters

n_i:Number of observations in the *i-th* cluster

n_{io}: Number of original observations in the *i-th* cluster

w_i: *1*

c: 0.5

- Completely indistinguishable when $U_c \to 1$, otherwise $\to 0.75$

Evaluation of identification disclosure risk

Disclosure risk measure using neighborhood-based approach adapted from Hu and Savitsky (2019)

$$S_r = \frac{1}{N}\sum_{i=1}^{N} I_i$$

S_r*:* Risk score

N: Total number of records

I_i: *1* if the *i-th* synthetic record lies in the 10-neighbors of the *i-th* original record, otherwise *0*

- Completely distinguishable when $S_r \to 0$, otherwise $\to 1$

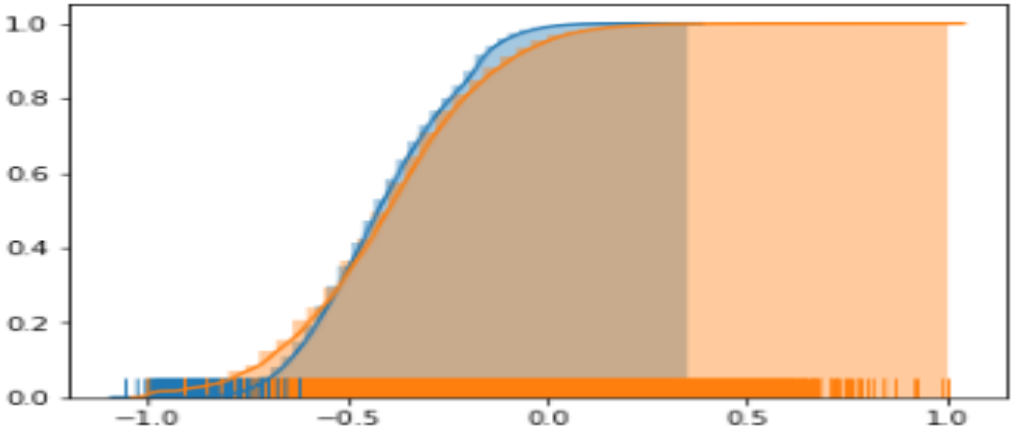
Pilot experiment

- A subset of 2012 Census of Agriculture dataset
  - One million records
  - No missing values
  - A subset of items selected
    - Total Land in Farms
    - Total Value Production
    - State (State id)
    - County (County id)
- Rescale to -1 and 1

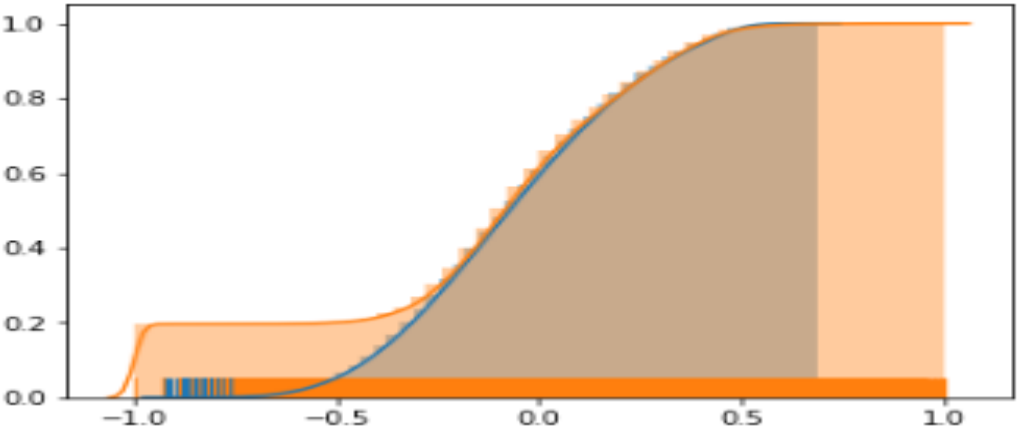
Pilot experiment - continued

- A GAN is trained on the subset of one million records dataset considering the loss function
  - In G-network
    - The cross entropy to penalize the output from G-network classified as synthetic by D-network
    - First, and second moments of the original distribution added to the loss function
  - In D-network
    - The cross entropy to penalize wrongly assigning the output from G-network to real, and real to synthetic
- Utility measures are calculated to evaluate the synthetic population
- Evaluation of identification disclosure risk measure is calculated

Result



Total Land in Farms



Total Value Production

- Original Distribution
- Synthetic Distribution

Result - continued

Propensity Score Measure: U_p = 0.97 → 1

Clustering Score Measure:

- *G* =200, U_c = 0.92 → 1

Risk Score Measure: S_r = 0.02 → 0

Conclusion

- GAN worked well for generating synthetic population for two continuous Census of Agriculture variables in terms of
  - Propensity score measure
  - Clustering score measure
- GAN failed to capture extreme values
- Identification disclosure risk of synthetic population is low

Future work

- Comprehensive experimentation
  - Multiple variables
  - Different numbers of clusters
  - Categorical, count, and skewed variables
- Further tuning of GAN's hyper-parameters
- Other measures of utility and of identification disclosure risks
- Adaptation to better fit distribution extremes

Selected references

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).
- Hu, J., & Savitsky, T. D. (2018). Bayesian data synthesis and disclosure risk quantification: An application to the Consumer Expenditure Surveys. *arXiv preprint arXiv:1809.10074*.
- Paiva, T., Chakraborty, A., Reiter, J. P., and Gelfand, A. E. (2014). Imputation of confidential data sets with spatial locations using disease mapping models. Statistics in Medicine 33, 1928–1945.
- Reiter, J. P. (2005a). Estimating risks of identification disclosure in microdata. Journal of the American Statistical Association 100, 1103–1112.
- Reiter, J. P. (2005b). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Journal of the Royal Statistical Society, Series A 168, 185–205.

Selected references

- Reiter, J. P. (2005c). Using cart to generate partially synthetic public use microdata. Journal of Official Statistics 21, 441–462.
- Rubin, D. B. (1993). Discussion statistical disclosure limitation. Journal of Official Statistics 9, 461–468
- Truta, T. M., Fotouhi, F., & Barth-Jones, D. (2003, July). Disclosure risk measures for microdata. In *15th International Conference on Scientific and Statistical Database Management, 2003.* (pp. 15-22). IEEE.
- Woo, M. J., Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, *1*(1).
- Drechsler, J. and Reiter, J. P. (2009), Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey, Journal of Official Statistics, 25, 589 - 603.

Questions?

ywei@niss.org