Approximate Nearest Neighbours Imputation

Maciej Beręsewicz
(Poznań Univeristy of Economics and Business,
Statistical Office in Poznań; the R guy)
Tomasz Hinc
(Poznań University of Economics and Business; the Python guy)

BigSurv, Barcelona

POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Table of contents

Motivation

1. Increasing non-contacts and non-response rates in sample surveys.
2. Variety of new, often large, data sources (e.g. big data) that are not error free (e.g. missing data).
3. Increasing need to impute (e.g. nearest neighbours) and link different sources (e.g. propensity matching, mass imputation).
4. Methods, based on distance matrices, used in official statistics are not suitable to deal with large, high dimensional, data sources.
5. On the other hand, industry developed variety of different high-performance open source libraries for neighbours search (e.g. `Annoy` by Spotify, `FAISS`, `pysparnn` by Facebook).
6. According to authors' knowledge there are no studies regarding impact of using approximate nearest neighbour algorithms on imputation estimators.

Motivation

1. Increasing non-contacts and non-response rates in sample surveys.
2. Variety of new, often large, data sources (e.g. big data) that are not error free (e.g. missing data).
3. Increasing need to impute (e.g. nearest neighbours) and link different sources (e.g. propensity matching, mass imputation).
4. Methods, based on distance matrices, used in official statistics are not suitable to deal with large, high dimensional, data sources.
5. On the other hand, industry developed variety of different high-performance open source libraries for neighbours search (e.g. `Annoy` by Spotify, `FAISS`, `pysparnn` by Facebook).
6. According to authors' knowledge there are no studies regarding impact of using approximate nearest neighbour algorithms on imputation estimators.

Motivation

1. Increasing non-contacts and non-response rates in sample surveys.
2. Variety of new, often large, data sources (e.g. big data) that are not error free (e.g. missing data).
3. Increasing need to impute (e.g. nearest neighbours) and link different sources (e.g. propensity matching, mass imputation).
4. Methods, based on distance matrices, used in official statistics are not suitable to deal with large, high dimensional, data sources.
5. On the other hand, industry developed variety of different high-performance open source libraries for neighbours search (e.g. `Annoy` by Spotify, `FAISS`, `pysparnn` by Facebook).
6. According to authors' knowledge there are no studies regarding impact of using approximate nearest neighbour algorithms on imputation estimators.

Motivation

1. Increasing non-contacts and non-response rates in sample surveys.
2. Variety of new, often large, data sources (e.g. big data) that are not error free (e.g. missing data).
3. Increasing need to impute (e.g. nearest neighbours) and link different sources (e.g. propensity matching, mass imputation).
4. Methods, based on distance matrices, used in official statistics are not suitable to deal with large, high dimensional, data sources.
5. On the other hand, industry developed variety of different high-performance open source libraries for neighbours search (e.g. `Annoy` by Spotify, `FAISS`, `pysparnn` by Facebook).
6. According to authors' knowledge there are no studies regarding impact of using approximate nearest neighbour algorithms on imputation estimators.

Motivation

1. Increasing non-contacts and non-response rates in sample surveys.
2. Variety of new, often large, data sources (e.g. big data) that are not error free (e.g. missing data).
3. Increasing need to impute (e.g. nearest neighbours) and link different sources (e.g. propensity matching, mass imputation).
4. Methods, based on distance matrices, used in official statistics are not suitable to deal with large, high dimensional, data sources.
5. On the other hand, industry developed variety of different high-performance open source libraries for neighbours search (e.g. `Annoy` by Spotify, `FAISS`, `pysparnn` by Facebook).
6. According to authors' knowledge there are no studies regarding impact of using approximate nearest neighbour algorithms on imputation estimators.

Motivation

1. Increasing non-contacts and non-response rates in sample surveys.
2. Variety of new, often large, data sources (e.g. big data) that are not error free (e.g. missing data).
3. Increasing need to impute (e.g. nearest neighbours) and link different sources (e.g. propensity matching, mass imputation).
4. Methods, based on distance matrices, used in official statistics are not suitable to deal with large, high dimensional, data sources.
5. On the other hand, industry developed variety of different high-performance open source libraries for neighbours search (e.g. `Annoy` by Spotify, `FAISS`, `pysparnn` by Facebook).
6. According to authors' knowledge there are no studies regarding impact of using approximate nearest neighbour algorithms on imputation estimators.

What are approximate nearest neighbours?

Given a finite dataset $X \subset R^d$ and real $\epsilon > 0$, $x^* \in X$ is an ϵ-approximate nearest neighbour of query $q \in R^d$, if $\text{dist}(q, x^*) \leq (1+\epsilon)\text{dist}(q, x)$ for all $x \in X$.

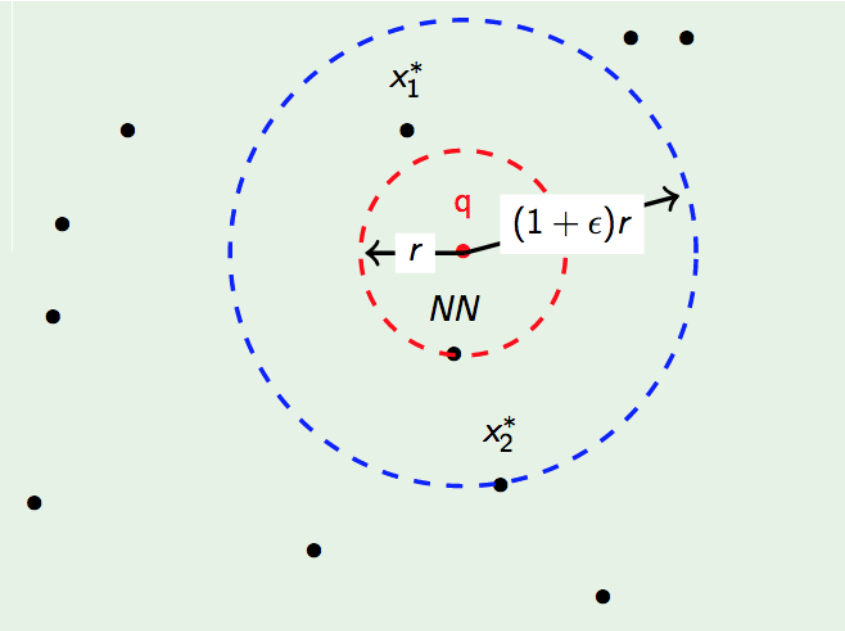

Figure 1: Illustration of exact and approximate near neighbour. Source: Avrithis, Emiris and Samaras (2016)

Why to bother with other methods to find NN?

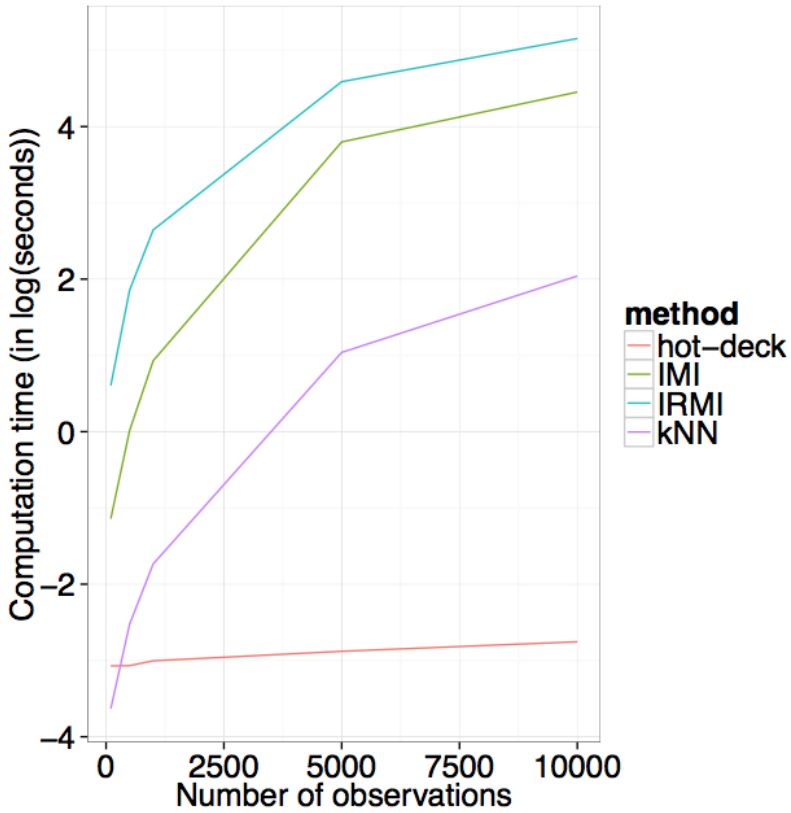

Figure 2: Comparison of computation times for four methods included in VIM package. Source: Kowarik and Templ (2016)

Research questions

1. How approximate nearest neighbours search is related to construction of imputation classes?
2. How approximate nearest neighbours imputation is related to other imputation methods used in official statistics such as exact nearest neighbours imputation, predictive mean matching or hot-deck imputation?
3. What is the performance of approximate nearest neighbours imputation in terms of bias and mean square error? Is there a difference between implementation of algorithms?

Literature review – imputation

- Imputation classes / cells
  - **Number of imputation classes and how they are created** – e.g. Little (1988), Haziza & Beaumont (2007).
  - **CART** – e.g. Toth & Eltinge (2011), Phipps & Toth (2012).
- k (exact) nearest neighbour imputation
  - **Theoretical basis of k-nearest imputation (inference)** – Chen & Shao (2000, 2001), Andridge & Little (2010), Yang & Kim (2017a),
  - **Matching on propensity scores (incl. predictive mean matching)** – Yang & Kim (2017a), Abadie & Imbens (2006, 2008, 2011, 2016),
  - **Mass imputation** – Kim & Wang (2018), Yang & Kim (2018), Rivers (2007).
- Other topics
  - **Record linkage and de-duplication** – cf. Chen, Shrivastava & Steorts (2018), Schnell (2015).

Literature review – the notation

- Let $\mathcal{F}_N = \{(\boldsymbol{x}_i, y_i, \delta_i) : i = 1, ..., N\}$ denote a finite population, where $\boldsymbol{x}_i$ is always observed, y_i has missing values and $\delta_i = \{0, 1\}$ is response indicator.
- We assume that $Pr(\delta = 1|x, y) = Pr(\delta = 1, x)$.
- Objective is to estimate $\mu = N^{-1} \sum_{i=1}^{N} g(y_i)$ for some known $g(.)$.
- The nearest neighbour imputation (NNI) estimator is given by

$$\hat{\mu}_{g,\mathrm{NNI}} = \frac{1}{N} \sum_{i \in A} \frac{1}{\pi_i} \left\{ \delta_i g\left(y_i\right) + \left(1 - \delta_i\right) g\left(y_{i(1)}\right) \right\}, \tag{1}$$

  where $y_{i(1)}$ is nearest neighbour that satisfies $d(x_{i(1)}, x_i) \leq d(x_j, x_i)$.
- The imputation estimator based on predictive mean matching (PMM) is given by

$$\hat{\mu}_{\mathrm{RMM}} = \frac{1}{N} \sum_{i \in A} \frac{1}{\pi_i} \left\{ \delta_i y_i + \left(1 - \delta_i\right) y_{i(1)} \right\}, \tag{2}$$

  where $y_{i(1)}$ is nearest neighbour that satisfies $d(\hat{m}_{i(1)}, \hat{m}_i) \leq d(\hat{m}_j, \hat{m}_i)$ where $\hat{m}_i = m(x_i, \beta)$.

Literature review – what we know about imputation by NN

Following, Abadie & Imbens (2006, 2016), Yang & Kim (2017a), we can study asymptotic properties of imputation estimator $\hat{\mu}_{g,NNI}$ by decomposing as follows

$$\sqrt{n}\,(\hat{\mu}_{g,NNI} - \mu_g) = D_N + B_N, \tag{3}$$

where D_N is variance and B_N is bias defined as follows

$$B_n = \frac{\sqrt{n}}{N} \sum_{i \in A} \frac{1}{\pi}(1 - \delta_i)\{\mu_g(x_{i(1)}) - \mu_g(x_i)\}, \tag{4}$$

where $\mu_g(x_{i(1)}) - \mu_g(x_i)$ is the matching discrepancy and B_n contributes to the asymptotic bias of the matching estimator.

- Abadie & Imbens (2006, 2016) showed that if the matching variable x is p-dimesional then $d(x_{i(1)}, x_i) = O_p(n^{-1/p})$.
- This means that if $p \geq 2$ the bias $B_N = O_p(n^{1/2-1/p}) \neq o_p(1)$.

Literature review – what we know about imputation by PMM

Following, Abadie & Imbens (2006, 2016), Yang & Kim (2017a), we can study asymptotic properties of predictive mean matching estimator $\hat{\mu}_{g,PMM}$ by decomposing as follows

$$\sqrt{n}\,(\hat{\mu}_{g,PMM} - \mu_g) = D_N(\beta) + B_N(\beta), \tag{5}$$

where $D_N(\beta)$ is variance and $B_N(\beta)$ is bias defined as follows

$$B_n = \frac{\sqrt{n}}{N} \sum_{i \in A} \frac{1}{\pi}(1 - \delta_i)\{m(x_{i(1)}; \beta) - m(x_i; \beta)\}, \tag{6}$$

where $m(x_{i(1)}; \beta) - m(x_i; \beta)$ is the matching discrepancy and B_n contributes to the asymptotic bias of the matching estimator.

- For PMM we use $m(x)$ to look for nearest neighbours, which is done scalar and hence $B_N = O_p(n^{1/2-1/1}) = o_p(1)$ which is asymptotically negligible.

Classification of ANN search

Classification of methods:

- Space partitioning methods – hierarchical partitioning methods; mainly based on creating trees (graphs).
- Locality Sensitive Hashing (LSH) methods.
- Filter-and-refine methods based on projection to a lower-dimensional space.
- Filtering methods based on permutations .
- Methods that construct a proximity graph.
- Miscellaneous methods.

Common steps for approximate nearest neighbour search:

1. Build an index of observations based on variables from input data and selected algorithm.
2. Query the index based on the same variables with selected radius (higher the radius, faster the search).

Idea of ANN – k-d trees

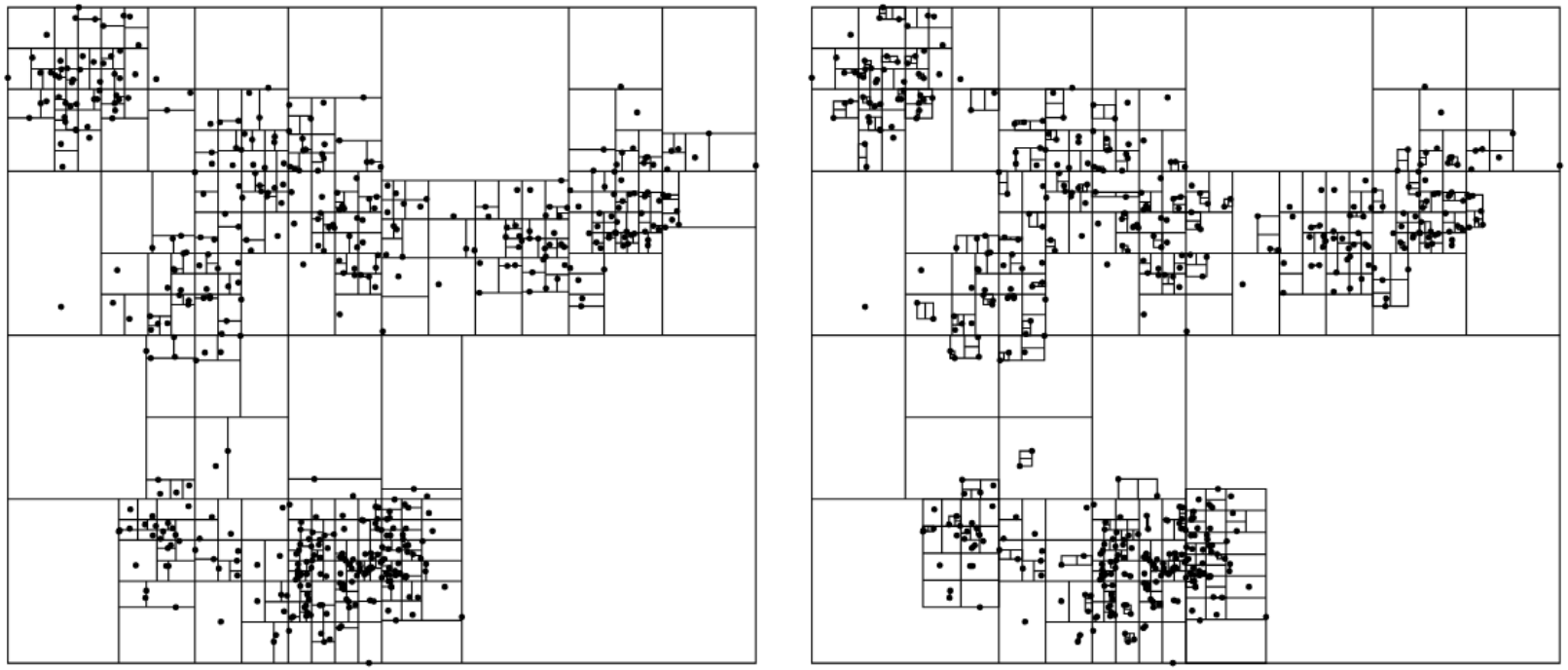

Figure 3: Kd-tree (left) and a box-decomposition tree (right). Source: Mount (2006)

Idea of ANN – hyperplanes (e.g. Annoy)

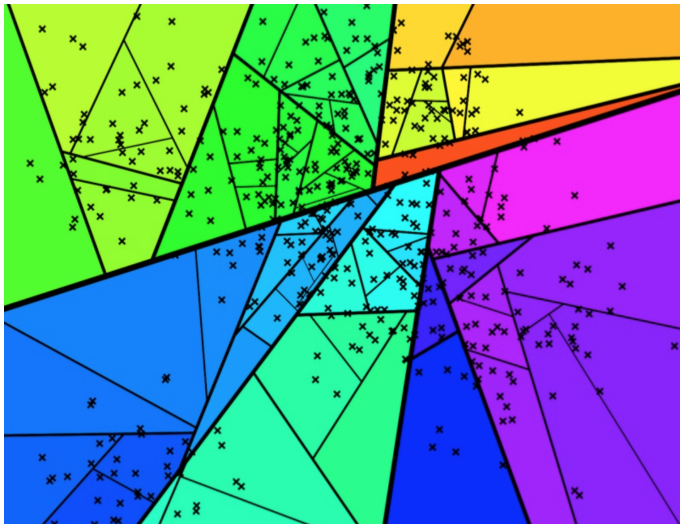

Figure 4: Annoy random hyperplanes used to divide the space. Source: Bernhardsson (2015)

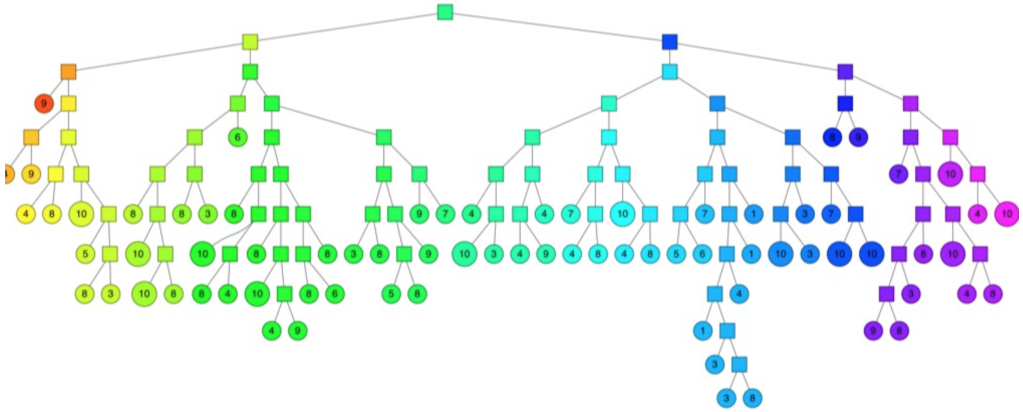

Figure 5: Binary trees for hyperplanes. Source: Bernhardsson (2015)

Idea of ANN – Locality Sensitive Hashing

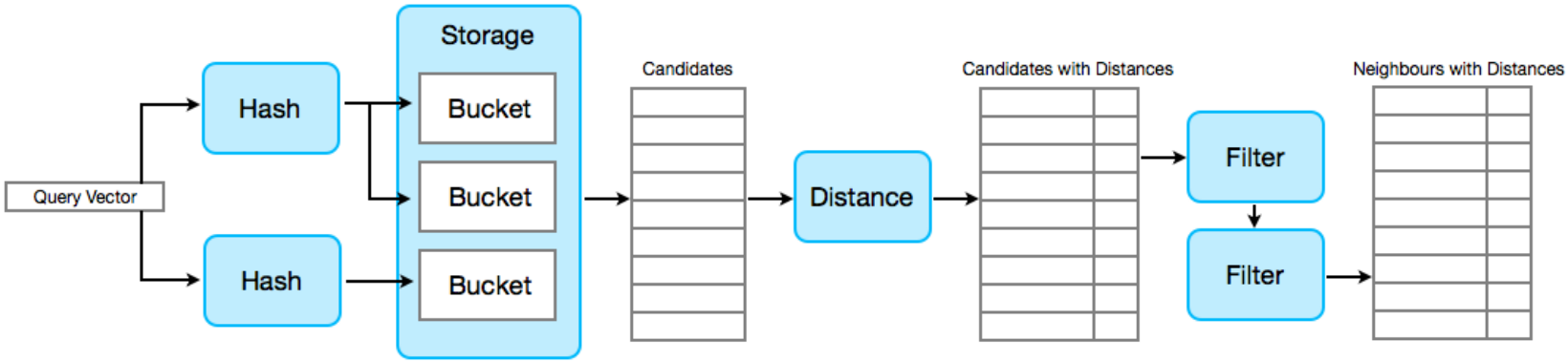

Figure 6: NearPy pipeline based on hashing data

Literature review – open-source software

- Imputation related
  - `yaImpute::ann` – it contains `ANN` (C++) library for exact searching of nearest neighbours (Crookston and Finley (2007)).
  - `miceExt::match_multivariate` – it uses `ANN` (C++) library for exact predictive mean matching (from `RANN` package).
- General use
  - `R` – `FNN`, `RANN`, `RANN.L1`, `nmslibR`, `RcppAnnoy` and many more,
  - `Python` – `nn2`, `Annoy`, `nmslib`, `faiss`, `sklearn.neighbors.NearestNeighbors`, `sklearn.neighbors.LSHForest` and many more.

Simulation studies – data mechanisms

For the simulation studies we generated population variables (cf. Yang & Kim 2017a,b) with sample sizes N = 10 000, 50 000 according to the following models

$$\begin{aligned} Y_1 &= -1 + X_1 + X_2 + \epsilon, \\ Y_2 &= -1.167 + X_1 + X_2 + (X_1 - 0.5)^2 + (X_2 - 0.5)^2 + \epsilon, \\ Y_3 &= -1.5 + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + \epsilon, \end{aligned} \tag{7}$$

where X_1, X_2, X_3 where generated independently from $U[0,1]$ and X_4, X_5, X_6, ϵ were generated independently $N(0,1)$. The response mechanism (75%) was generated according to

$$\log\left(\frac{\rho}{1-\rho}\right) = 0.2 + X_1 + X_2. \tag{8}$$

We did not consider non-ignorable non-response as hot-deck (incl. nearest neighbour) imputation are not suited for this case.

(Limited) simulation study – settings

For the simulation studies we considered:

- The census case – using only a response mechanism,

In the simulation study we focused on **bias** and **MSE** of exact and approximate nearest neighbours for the following settings:

- matching based on $m()$ (i.e. predictive mean matching),
- matching based on X_1, X_2.

For both cases we consider Euclidean distance (i.e. L2 norm) and only one neighbour (k=1).

Simulation study – settings

1. **Different algorithms to search donors and impute missing data**:
   - `Hot-deck` imputation with no imputation classes (**HD**), 5 classes based on quantiles.
   - `K-D tree` exact nearest neighbours (**KD-ENN**),
   - `K-D tree` approximate nearest neighbours (**KD-ANN**),
   - `Annoy` approximate nearest neighbours (**Annoy**),
2. For the simulation study we used R 3.5.1 (Microsoft R Open) and packages:
   - `RcppAnnoy 0.0.10`,
   - `yaImpute 1.0.29`.

Simulation study – measures of quality

- Monte Carlo Absolut Relative Bias (RB) given by

$$\mathrm{ARB}(\bar{y}_I) = \left|\frac{E_{MC}(\bar{y}_I) - \bar{Y}}{\bar{Y}} \times 100\%\right|, \tag{9}$$

where $E_{MC} = R^{-1}\sum_{r=1}^{R} \bar{y}_I^{(r)}$ and $\bar{y}_I^{(r)}$ is the imputed estimator in r-th sample, $r = 1, .., R$

- Monte Carlo Relative Absolute Root Mean Square Error (RRMSE) given by

$$\mathrm{ARRMSE}(\bar{y}_I) = \left|\frac{\sqrt{\mathrm{MSE}_{MC}(\bar{y}_I)}}{\bar{Y}} \times 100\%\right|, \tag{10}$$

where $\mathrm{MSE}_{MC}(\bar{y}_I) = R^{-1}\sum_{r=1}^{R}\left(\bar{y}_I^{(r)} - \bar{Y}\right)^2$

We set $R = 500$.

Simulation study – the census case (N=10 000)

Imputation based on $m()$ i.e. predictive mean matching using hot-deck with/without imputation classess, exact and approximate NN and Annoy.

Method	Var	ARB	ARRMSE	Method	Var	ARB	ARRMSE
Hot-deck	Y_1	385.03	393.60	ANN	Y_1	1.90	70.88
(no classes)	Y_2	367.90	376.95	(eps=2)	Y_2	1.61	77.23
	Y_3	387.25	414.65		Y_3	9.70	77.15
Hot-deck	Y_1	31.66	82.05	ANN	Y_1	3.25	70.23
(5 classes)	Y_2	24.82	79.22	(eps=5)	Y_2	0.95	77.26
	Y_3	52.52	100.03		Y_3	8.73	77.57
ANN	Y_1	1.76	**71.24**	Annoy	Y_1	386.43	392.98
(exact)	Y_2	1.97	**77.25**	(trees=1)	Y_2	370.28	376.50
	Y_3	9.72	**76.52**		Y_3	383.48	403.61
ANN	Y_1	**1.74**	70.85	Annoy	Y_1	386.42	392.97
(eps=1)	Y_2	**1.47**	77.46	(trees=5)	Y_2	370.23	376.46
	Y_3	**9.47**	76.62		Y_3	383.49	403.62

Simulation study – the census case (N=10 000)

Imputation based on X_1, X_2 using exact and approximate NN and Annoy.

Method	Var	ARB	ARRMSE
KD-ENN (exact)	Y_1	**21.55**	**74.28**
	Y_2	**26.60**	**77.48**
	Y_3	**45.08**	**148.40**
KD-ANN (eps=1)	Y_1	35.10	78.49
	Y_2	40.58	84.16
	Y_3	57.20	150.09
KD-ANN (eps=2)	Y_1	35.71	79.39
	Y_2	43.22	85.24
	Y_3	72.85	158.67

Method	Var	ARB	ARRMSE
KD-ANN (eps=5)	Y_1	**13.60**	**70.79**
	Y_2	**22.03**	**78.00**
	Y_3	51.67	150.42
Annoy (trees=1)	Y_1	**14.77**	**73.28**
	Y_2	**20.36**	**77.33**
	Y_3	**35.34**	**147.77**
Annoy (trees=100)	Y_1	22.53	74.73
	Y_2	27.25	77.80
	Y_3	44.44	148.79

Conclusion

Based on our (limited) simulation studies the following conclusions may be stated:

- Approximate nearest neighbour creates imputation classes and due to that performs better than hot-deck imputation.
- Imputation classes are created on a dataset without missing data and thus we avoid problems in creating cells (e.g no donors).
- Annoy performs poorly in 1d space while for higher dimensions is better.
- Looking for approximate nearest neighbours in close range (eps=1,2) may lead to decrease in ARB and ARRMSE.

General remarks regarding the topic

- Each method has many different parameters and can be tuned.
- In most cases methods are based on different heuristic algorithms and implementations can be very different.

References I

- Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects, Econometrica 74: 235–267.
- Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators, Econometrica 76: 1537–1557.
- Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects, Journal of Business & Economic Statistics 29: 1–11.
- Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score, Econometrica 84: 781– 807.
- Avrithis, Emiris and Samaras (2016)
- Avarikioti, G., Emiris, I. Z., Psarros, I., & Samaras, G. (2016). Practical linear-space Approximate Near Neighbors in high dimension. arXiv preprint arXiv:1612.07405.
- Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. International statistical review, 78(1), 40-64. Bernhardsson, Erik (2015), Approximate nearest neighbor methods and vector models – NYC ML meetup, https://www.slideshare.net/erikbern/approximate-nearest-neighbor-methods-and-vector-models-nyc-ml-meetup
- Chen, B., Shrivastava, A., & Steorts, R. C. (2018). Unique entity estimation with application to the Syrian conflict. The Annals of Applied Statistics, 12(2), 1039-1067.
- Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. Journal of Official statistics, 16(2), 113.

References II

- Chen, J., & Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. Journal of the American Statistical Association, 96(453), 260-269.
- Crookston, Nicholas L.; Finley, Andrew O. 2007. yaImpute: An R Package for k-NN Imputation. Journal of Statistical Software. 23(10):1-16.
- Haziza, D., & Beaumont, J. F. (2007). On the construction of imputation classes in surveys. International Statistical Review, 75(1), 25-43.
- Kim, J. K., & Wang, Z. (2018). Sampling Techniques for Big Data Analysis. International Statistical Review.
- Kowarik, A. & Templ M. (2016). Imputation with the R Package VIM. Journal of Statistical Software, 74(7), 1-16. doi:10.18637/jss.v074.i07
- Little, R. J. (1988). Missing-data adjustments in large surveys. Journal of Business & Economic Statistics, 6(3), 287-296.
- Mount, D. M. (2006). ANN programming manual. http link: http://www. cs. umd. edu/ mount/ANN/Files/1.1, 1. Phipps, P., & Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. The Annals of Applied
- Statistics, 6(2), 772-794.
- Rivers, D. (2007). Sampling for web surveys. In ASA Proceedings of the Section on Survey Research Methods. American Statistical Association, Alexandria, VA, pp. 4127–4134.
- Schnell, R. (2015). Privacy-preserving record linkage.

References III

- Toth, D., & Eltinge, J. L. (2011). Building consistent regression trees from complex sample data. Journal of the American Statistical Association, 106(496), 1626-1636.
- Yang, S., & Kim, J. K. (2017a). Nearest neighbor imputation for general parameter estimation in survey sampling. arXiv preprint arXiv:1707.00974.
- Yang, S., & Kim, J. K. (2017b). Predictive mean matching imputation in survey sampling. arXiv preprint arXiv:1703.10256.
- Yang, S., & Kim, J. K. (2018). Integration of survey data and big observational data for finite population inference using mass imputation. arXiv preprint arXiv:1807.02817.