Probabilistic record linkage through recursive partitioning without training data

Abel Dasylva (Statistics Canada) and Winfield Chen (Simon Fraser University)

April 6, 2022

The content of this presentation represents the authors' opinions and not necessarily those of Statistics Canada.

Éclairer grâce aux données, pour bâtir un Canada meilleur

Outline

1. Motivation
2. Background
3. Selecting the features
4. Making optimal decisions
5. Practical considerations
6. Example
7. Conclusion

1. Motivation

- At Statistics Canada, many important studies use the probabilistic method to link the census to another source such as
  - tax records (Satistics Canada, 2017, Pinault et al., 2016) or
  - mortality records (Wilkins et al., 2008).

1. Motivation (cont'd)

- Such linkages are implemented with G-LINK, where there is a very large choice of comparison functions (aka features) to select from.
- Each variable may be compared at many levels of agreement including
  - full agreement, e.g. same given name, or
  - partial agreements such as phonetic, typo, jaro-winkler string similarity, etc.
- G-LINK also supports matrix comparisons and conditional comparisons to jointly compare many variables at different agreement levels.

1. Motivation (cont'd)

- The result is a bewildering array of options to manually choose from, even with the usual social variables, i.e. given name, surname, sex, birth date, address.
- A process that is challenging and labour intensive when there is no training data (typically the case).
- Besides the result may be far from optimal.

1. Motivation (cont'd)

- How to automate and optimize the selection of the features without training data?
  - Reduce the costs.
  - Increase the timeliness.
  - Increase the precision.
- Can open source packages (in R or Python) help?

2. Background

- Identify the record pairs, where the records come from the same unit, e.g. a person, household or business.
- These pairs are called matched while the remaining pairs are called unmatched.
- Errors occur when the linkage decisions are based on quasi-identifiers, such as names and dates.

2. Background (cont'd)

- The steps of the probabilistic method (Fellegi and Sunter, 1969)

1. Apply blocking criteria.	2. Compare the records in each pair.	3. Assign a similarity to each pair.	4. Link the pairs with sufficient similarity.
Get a subset of the Cartesian product when linking two files.	Generate the features.	The similarity is a function of the features.	Use a similarity threshold.

2. Background (cont'd)

- The similarity of a pair is measured by the linkage weight

$$w = 10log_2\left(\frac{P(observed\ features|matched)}{P(observed\ features|unmatched)}\right)$$

- The weight threshold is set according to the error target.
- An error is a false negative or a false positive.
  - A false negative (FN) is failing to link records from the same unit.
  - A false positive (FP) is linking records from different units.

2. Background (cont'd)

- The probabilistic method makes optimal trade-offs between the false negatives and the false positives (Fellegi and Sunter, 1969).
- Limitations and challenges
  - How to select the features?
  - How to estimate the decision parameters (i.e. the weights and the thresholds)?
  - How to estimate the errors?
  - No truth deck or training data

2. Background (cont'd)

- Estimate the errors by modeling the number of links adjacent to a record (Dasylva and Goussanou, 2020, 2021) when linking a file to a census s.t.
  a. the census is complete (i.e. no undercoverage) with N records,
  b. both sources are free of duplicate records and
  c. the decision to link two records involves no other record.
- These conditions are nearly met in applications, such as the linkage of tax records to the census (Statistics Canada, 2017) or the study of the mortality of a census cohort (Blakely and Salmond, 2002).

2. Background (cont'd)

- The parameters of the error model are of the form $\left[\left(\alpha_g, p_g, \lambda_g\right)\right]_{1 \leq g \leq G}$ where
  - $\bar{p} = \sum_{g=1}^{G} \alpha_g p_g$ is the recall, i.e. the proportion of matched pairs that are linked.
  - $\bar{p}/\left(\bar{p} + \bar{\lambda}\right)$ is the precision, i.e. the proportion of linked pairs that are matched, with $\bar{\lambda} = \sum_{g=1}^{G} \alpha_g \lambda_g$.
  - $w = 10log_2\left((N-1)\bar{p}/\bar{\lambda}\right)$ is the corresponding probabilistic linkage weight.
- The estimates require no training data and account for all the interactions.

3. Selecting the features

- The manual optimization of the features is impractical.
- Automate by recursive partitioning (Hastie et al., 2001, chap 9.2).
  - It is intuitive and naturally accounts for the features interactions.
  - Aim for a homogeneous conditional match probability within each leaf.
  - Select the features that are good predictors of the match status.
  - A feature is selected if it is used to split some node.

3. Selecting the features (cont'd)

- How to build the tree without any training data?
- Link a file to a complete census, where each source is without duplicates.
- For the pair including the file record i and the census record j, define
  - the initial features $X_{ij} = \left[X_{ij}^{(k)}\right]_{1 \le k \le K}$ representing different choices of comparison functions and
  - the match status Y_{ij} indicating whether the pair is matched.

3. Selecting the features (cont'd)

- When the node impurity is based on Gini's index, the impurity of a branch R is given by $\hat{q}(1-\hat{q})$ (Hastie et al., 2001, chap. 9.2) where

$$\hat{q} = \left(\sum_{(i,j)} I(X_{ij} \in R) Y_{ij}\right) \Big/ \left(\sum_{(i,j)} I(X_{ij} \in R)\right)$$

  is the <u>precision</u> when linking all the pairs of the branch.

- Estimate the precision $\hat{q}$ by $\bar{p}/(\bar{p}+\bar{\lambda})$ with the model, i.e. without any training data, while accounting for all the features interactions.

3. Selecting the features (cont'd)

- Alternative procedure for building the tree without training data
  - Use the k-means to classify the pairs and build the tree using the assigned labels (Elfeky et al., 2003). Implemented in the RecordLinkage R package.
  - Training data is still required to estimate the linkage accuracy.
- Model-based trees by Loh (2002) or Zeileis et al. (2008)
  - They apply to the analytical variables.
  - They incorporate a model but still require some training data.

4. Making optimal linkage decisions

- Applying the probabilistic method to the tree leaves.
- For each leaf, estimate the probabilistic weight by linking all the pairs therein and estimating the leaf weight with the model.
- Link the pairs where the leaf weight is no less than a single threshold (i.e. no grey zone) based on a recall target.
- Estimate the overall precision and recall (expected to be close to the target) with the model.

5. Practical considerations

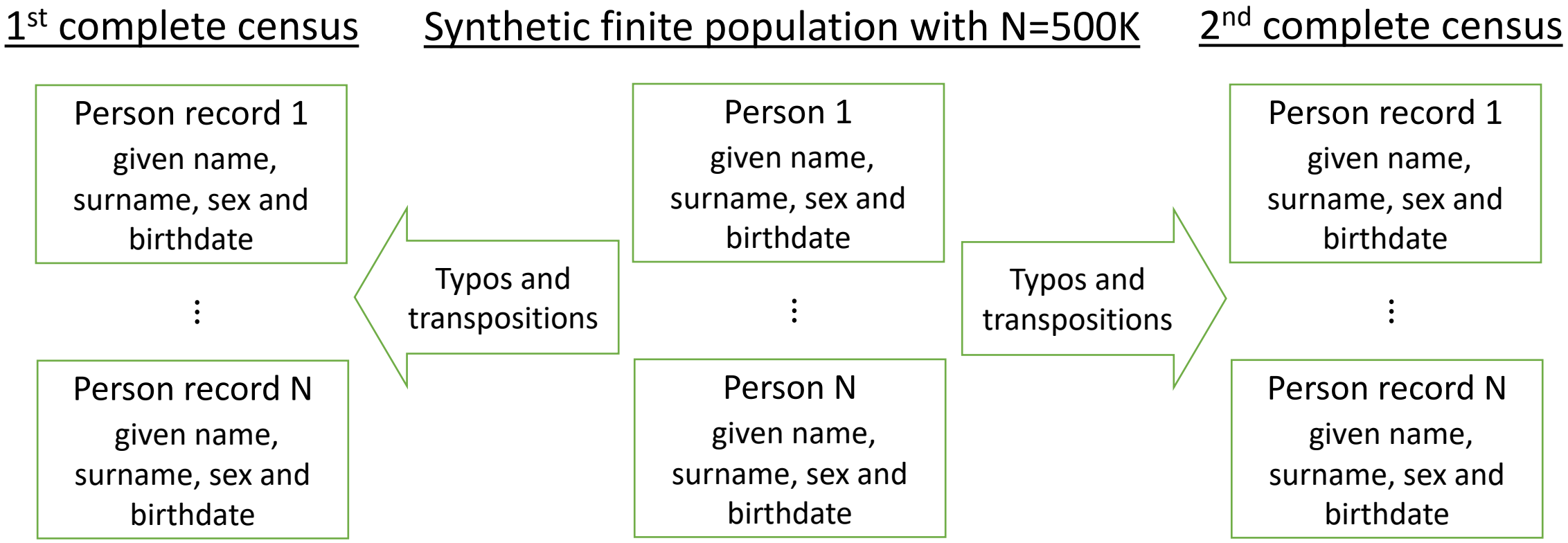
- Implement the solution with the R package Rpart and a custom splitting function (Therneau, 2019).
- For reliable model estimates, do not split a node if the estimated recall (i.e. $\bar{p}$) is too small, e.g. less than 0.1. This threshold is akin to a complexity parameter and it controls the number of leaves.
- No pruning.

5. Practical considerations (cont'd)

- For simplicity, consider binary features, where each is possibly based on many variables.
- The features may overlap in terms of the underlying variables.
- Due to memory constraints, the implementation is currently limited to a few millions pairs, depending on the number of features.

6. Experiment

- Use 2000 and 2010 US census data (US Census Bureau, 2016, 2020, 2021).



6. Experiment (cont'd)

- Blocking
  - The number indicates the corresponding census.
  - DD, MM and YYY are the day, month and year of birth, respectively.

YYYY1=YYYY2

AND

SOUNDEX(SURNAME 1)=SOUNDEX(SURNAME 2)
OR
SOUNDEX(SURNAME 1)=SOUNDEX(GIVEN NAME 2)

AND

DD1=DD2
OR
MM1=MM2
OR
DD1=MM2
OR
MM1=DD2

6. Experiment (cont'd)

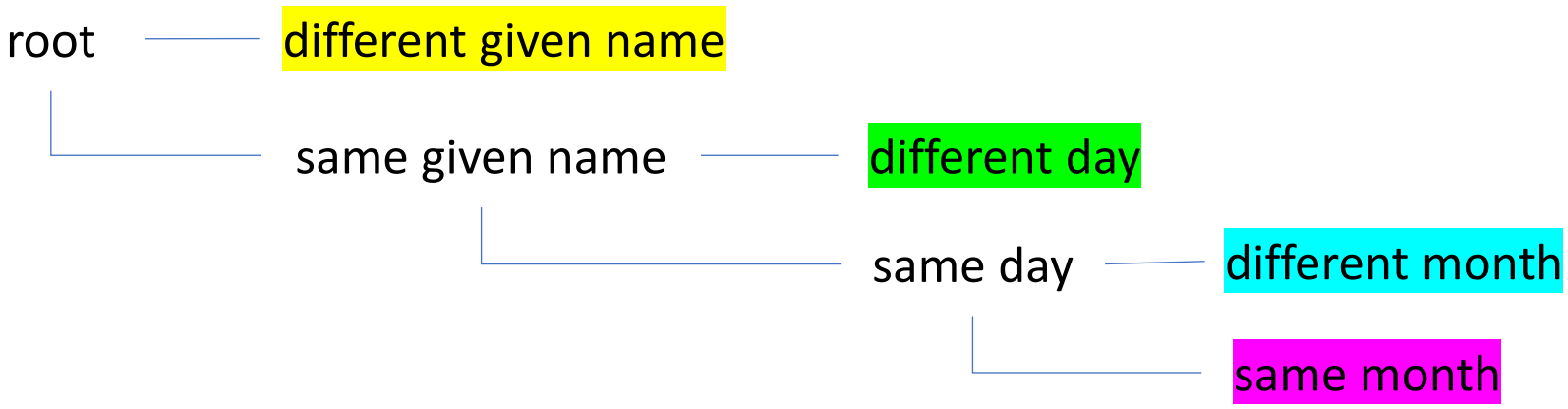
- Seven binary features based on exact comparisons

record from the 2nd census

record from the 1st census	sex	day	month	given name	surname
sex	X				
day		X	X		
month			X		
given name				X	X
surname					X

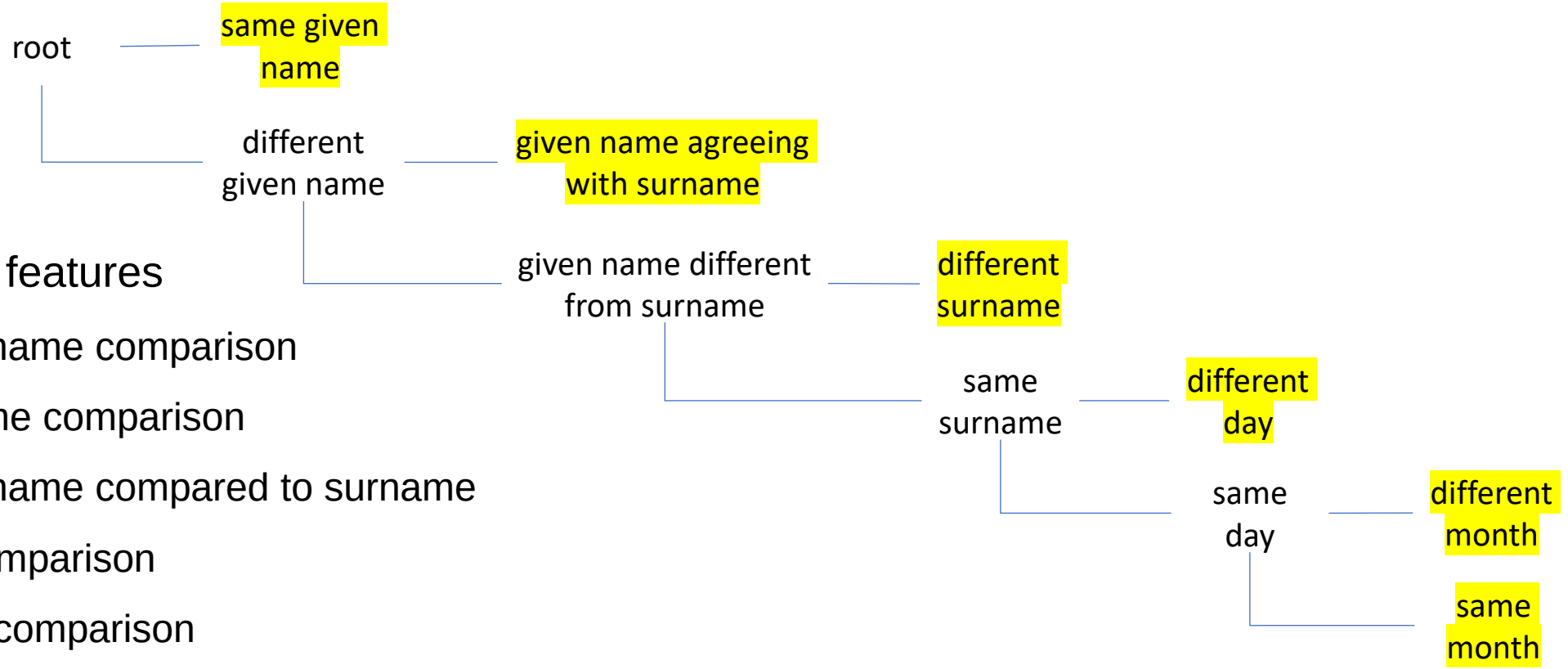
6. Experiment (cont'd)

- Model-based tree where the following features are selected.
  - Given name comparison
  - Day comparison
  - Month comparison



6. Experiment (cont'd)

- Compare with the tree that exploits all the truth.



Selected features

- Given name comparison
- Surname comparison
- Given name compared to surname
- Day comparison
- Month comparison

6. Experiment (cont'd)

- Naive estimates for the model-based tree
  - For each leaf, consider that we link the pairs therein and apply the model.
  - The estimates may be far from the actual values.

	Recall		Precision	
Leaf	*Estimate*	*Actual*	*Estimate*	*Actual*
diff. given name	0.0078	0.1402	0.0031	0.0495
same given name but diff. day	0.0110	0.1173	0.0049	0.5268
same given name and day but diff. month	0.0488	0.0574	0.5615	0.6604
same given name, day and month	0.5339	0.5342	0.9949	0.9954

6. Experiment (cont'd)

- Refined estimates through differencing
  - For each leaf except the last one, compare the estimates for the last leaf with those when linking the pairs that are within the said leaf or the last leaf.

	Recall		Precision		
Leaf	*Estimate*	*Actual*	*Estimate*	*Actual*	Estimated weight
diff. given name	0.1179	0.1402	0.0468	0.0495	146
same given name but diff. day	0.1197	0.1173	0.5370	0.5268	191
same given name and day but diff. month	0.0598	0.0574	0.8678	0.6604	216
same given name, day and month	0.5339	0.5342	0.9949	0.9954	265

6. Experiment (cont'd)

- Link the two censuses subject to a recall target of 0.7.
  - Link the pairs with weight no less than 216.
  - Link the pairs with weight 191 with the probability (0.7-0.5937)/0.1197=0.89.
  - Do not link the remaining pairs.

Leaf	Estimated recall	Cumulative estimated recall	Estimated weight
same given name, day and month	0.5339	0.5339	265 (L)
same given name and day but diff. month	0.0598	0.5937	216 (L)
same given name but diff. day	0.1197	0.7134	191 (L)
diff. given name	0.1179	0.8313	146 (NL)

decreasing weight

6. Experiment (cont'd)

- Link the two censuses subject to a recall target of 0.7.
  - Link the pairs with weight no less than 216.
  - Link the pairs with weight 191 with the probability (0.7- 0.5937)/0.1197=0.89.
  - Do not link the remaining pairs.
- Estimate the overall error.

	Recall	Precision
Actual	0.6958	0.8473
Model estimate	0.6850	0.8344

7. Conclusion

- The proposed methodology automates the selection of features when applying the probabilistic method without training data.
  - Also estimate the overall linkage accuracy without training data.
- In future work
  - Evaluate the methodology on actual data.
  - Refine the parameter estimation during the tree construction.
  - Extend the methodology for bigger datasets and more general linkages.
  - Investigate how a similar solution may be developed for blocking.

MERCI! / THANK YOU!

For more information,

abel.dasylva@statcan.gc.ca

Please visit www.statcan.gc.ca

#StatCan100

References

Blakely, T., and Salmond, C. (2002). "Probabilistic record linkage and a method to calculate the positive predicted value", *Journal of Epidemiology*, 31, 1246-1252.

Comenetez, J. (2016). “Demographic Aspects of Surnames-2020 Census”, https://www2.census.gov/topics/genealogy/2010surnames/surnames.pdf.

Dasylva, A., and Goussanou, A. (2020). “Estimating linkage errors under regularity conditions”, *in Proceedings of the Survey Methods Section*, American Statistical Association.

Dasylva, A., and Goussanou, A. (2021). “Estimating the false negatives due to blocking in record linkage”, *Survey Methodology*, 47, 299-311.

References (cont'd)

Elfeky, Mohamed G.; Verykios, Vassilios S.; Elmagarmid, Ahmed K.; Ghanem, Thanaa M.; and Huwait, Ahmed R. (2003), "Record Linkage: A Machine Learning Approach, A Toolbox, and a Digital Government Web Service", Department of Computer Science Technical Reports. Paper 1573.

Fellegi, I., and Sunter, A. (1969). "A theory of record linkage", *Journal of the American Statistical Association*, 64, 1183-1210.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.

Loh, W.Y. (2002). "Regression Trees With Unbiased Variable Selection and Interaction Detection", *Statistica Sinica*, 12, 361-386.

References (cont'd)

Pinault, L., Finès, P., Labrecque-Synnott, F., Saidi, A., and Tjekpema, M. (2016). "The 2001 Canadian Census-Tax-Mortality Cohort: A 10-Year Follow-up", *Analytical Studies: Methods and References No. 003 (Catalogue 11-633-X)*, Ottawa: Statistics Canada. Available at https://www150.statcan.gc.ca/n1/pub/11-633-x/11-633-x2016003-eng.htm.

Statistics Canada (2017a). 2016 census of population income reference guide. (98-500-X2016004)

Therneau, T. (2019). User written splitting functions for RPART, available at https://cran.r-project.org/web/packages/rpart/vignettes/usercode.pdf.

US Census Bureau (2016). "File B: Surnames Occurring 100 or more times", https://www2.census.gov/topics/genealogy/2010surnames/names.zip.

References (cont'd)

US Census Bureau (2020). "Annual State Resident Population Estimates for 6 Race Groups (5 Race Alone Groups and Two or More Races) by Age, Sex, and Hispanic Origin: April 1, 2010 to July 1, 2019", https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/asrh/sc-est2019-alldata6.csv.

US Census Bureau (2021, Oct. 8). Frequently Occurring Surnames from the Census 2000. https://www.census.gov/topics/population/genealogy/data/2000_surnames.html.

Wilkins, R., Tjepkema, M., Mustard, C., and Choinière, R. (2008). "The Canadian census mortality follow-up study, 1991 through 2001", *Health Reports*, 19, 25-43. Available at https://www150.statcan.gc.ca/n1/en/pub/82-003-x/2008003/article/10681-eng.pdf?st=Iq0FFyo0.

References (cont'd)

Wilkins, R., Tjepkema, M., Mustard, C., and Choinière, R. (2008). "The Canadian census mortality follow-up study, 1991 through 2001", *Health Reports*, 19, 25-43. Available at https://www150.statcan.gc.ca/n1/en/pub/82-003-x/2008003/article/10681-eng.pdf?st=Iq0FFyo0.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). "Model-based recursive partitioning", *Journal of Computational and Graphical Statistics*, 17, 492-514.