The K nearest neighbor algorithm for imputation of missing longitudinal prenatal alcohol data

Ayesha Sania (as4823@cumc.columbia.edu)
Columbia University Medical Center
Nicolò Pini
Columbia University Medical Center
Morgan E. Nelson
Avera Health
Michael M. Myers
Columbia University Medical Center
Lauren C. Shuffrey
Columbia University Medical Center
Maristella Lucchini
Columbia University Medical Center
Amy J. Elliott
Avera Health
Hein J. Odendaal
Stellenbosch University Faculty of Medicine and Health Sciences
William P. Fifer
Columbia University Medical Center

Research article

Keywords: K Nearest Neighbor, Machine Learning, Data Missingness, Data Imputation, Prenatal Alcohol Consumption, Longitudinal Alcohol Consumption

Posted Date: June 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-32456/v1

Abstract

Background — Missing data are a source of bias in many epidemiologic studies. This is problematic in alcohol research where data missingness may not be random as they depend on patterns of drinking behavior.

Methods — The Safe Passage Study was a prospective investigation of prenatal alcohol consumption and fetal/infant outcomes (n=11,083). Daily alcohol consumption for the last reported drinking day and 30 days prior was recorded using the Timeline Followback method. Of 3.2 million person-days, data were missing for 0.36 million. We imputed missing exposure data using a machine learning algorithm; "K Nearest Neighbor" (K-NN). K-NN imputes missing values for a participant using data of other participants closest to it. Since participants with no missing days may not be comparable to those with missing data, segments from those with complete and incomplete data were included as a reference. Imputed values were weighted for the distances from nearest neighbors and matched for day of week. We validated our approach by randomly deleting non-missing data for 5-15 consecutive days.

Results — We found that data from 5 nearest neighbors (i.e. K=5) and segments of 55 days provided imputed values with least imputation error. After deleting data segments from a first trimester data set with no missing days, there was no difference between actual and predicted values for 64% of deleted segments. For 31% of the segments, imputed data were within +/-1 drink/day of the actual.

Conclusions — K-NN can be used to impute missing data in longitudinal studies of alcohol use during pregnancy with high accuracy.

Background:

Accurate assessment timing, frequency, and quantity of prenatal alcohol exposure (PAE) in longitudinal research studies is necessary for obtaining unbiased assessments of the effects on fetal and infant outcomes. Despite the importance from a public health point of view, there are currently no robust biomarkers for assessing timing and amount of alcohol exposure during pregnancy. Thus, we often remain reliant on maternal self-report of intake. Aside from issues associated with the accuracy of self-report, there are other methodological challenges in measuring alcohol exposure in longitudinal studies[1, 2]. Recording daily intake, while providing a temporally complete set of values, involves significant participant burden and is likely to impact consumption behavior [3]. As a consequence, in many studies, alcohol consumption data are sampled at various times throughout pregnancy[4]. However, even when data for the specific collection time-points are complete, there is frequently missing information about intake during the intervals between samples. Addressing this missing data problem is critical when the exposure metrics of interest are both timing and amount in pregnancy [5].

The impact of missing data on the validity of estimates largely depends on the reasons data is missing [6]. For example, pregnant women of low socioeconomic status (SES) are more likely to access antenatal studies, and, therefore, have more missing data early in

pregnancy [7]. This is problematic as SES is an important determinant of drinking behavior during pregnancy [8]. In addition, women often modify their consumption behavior following pregnancy recognition, which can happen at varying times during the first months of pregnancy. While some women stop or reduce drinking immediately upon pregnancy recognition, some heavy drinkers continue to binge in the first trimester or continue heavy drinking throughout the pregnancy [5]. The accuracy of measures irrespective of the presence of missing data, such as the number of drinks consumed only on drinking days, may also provide biased overall estimates depending on when participants are interviewed. Accordingly, new approaches for managing the missing data problem are needed.

The Safe Passage Study conducted by the Prenatal Alcohol and SIDS and Stillbirth Network (PASS) was a prospective investigation of effects of alcohol exposure on multiple fetal and infant outcomes in Cape Town, South Africa and the Northern Plains, USA[9]. In this study, alcohol data were collected using a modification of the Timeline Followback Method (TLFB)[10], in which mothers recorded drinking data on their last known drinking day and then, for the 30 days prior. While this method was deemed the best self-report system available, the approach necessarily generated a variable amount of missing data.

Here, our goal was to impute the drinking values on missing days using a machine learning algorithm called k-nearest neighbor (*K*-NN). *K*-NN imputes missing values using pattern recognition without any distributional assumption about the underlying data[11]. The *K*-NN algorithm has been used in imputation of missing data in several research studies in the healthcare field[12, 13]. In this paper, we provide the methodological details of the specific application of the *K*-NN method for PASS exposure data and the validation of these results.

Methods:

The Safe Passage Study:

PASS was a prospective study of a cohort of pregnant women and their infants evaluating the role of prenatal alcohol exposure on incidence of adverse pregnancy outcomes including stillbirth, sudden infant death syndrome (SIDS), and fetal alcohol spectrum disorders (FASDs) of the surviving children. Between August 2007 and January 2015, 11,892 pregnant women were enrolled from antenatal clinics in Northern Plains, USA and Cape Town, South Africa. Women were eligible to participate in the study if they were pregnant with one or two fetuses, aged 16 years or older, were at gestational age 6 weeks or later at recruitment and spoke English or Afrikaans. Women were followed throughout the pregnancy and 1 year postnatally. Data on socioeconomic status, demographic, obstetric and medical history, periconceptional drinking and smoking were collected at the enrollment interview. Information on subsequent drinking during pregnancy was updated in study visits following enrollment. Written informed consent was obtained from all participants. Ethical approval was obtained from Stellenbosch University, Sanford Health, the Indian Health Service and from participating Tribal Nations.

Alcohol data collection method and missing data

Alcohol exposure data were collected using a modified validated TLFB [10], which required participants to report details of their drinking on each day ±15 days from the last menstrual period (LMP) and, at each study visit, the thirty days prior to the last known drinking day. Data were collected on the types of drinks, number of drinks, size of the containers, amount of ice in the drink, how many people drinks were shared with, and duration of the drinking episodes. These data were then used to estimate the total amount of alcohol consumed and number of standard drinks on each reported drinking day [14]. Data on drinking were collected during 1-4 prenatal study visits and 1 visit postpartum.

Due to the nature of the modified TLFB data collection design, the number of days with missing data varied by participant as a function of the time of enrollment and number of subsequent visits. The number of days with missing drinking information also varied for each participant depending on the recentness of their drinking. Figure 1 shows examples of how such variation emerged during the time period between LMP and the recruitment visit depending on when the last drinking day occurred. Participants who did not drink, or whose last drinking day was prior to their LMP had no missing data (Figure 1, panel a). Participants who drank but quit drinking within 30 days of the last collection period, had less or no missing data (Figure 1, panel b). Participants who continued to drink, and who reported drinking information 30 days closest to the interview date, had missing information prior to the 30-day period of reported drinking (Figure 1, panel c). In this example, if Subject Z drank often, and possibly at a higher volume, she would have a greater number of missing days than women who drink less often. Thus, a summation of drinks over the days will reflect less than the actual consumption and analysis using this exposure metric will be biased.

The KNN algorithm:

k-NN is a non-parametric machine learning algorithm which can be utilized to impute missing drinking information of a subject based on the information provided by other observations in a given database. Figure 2 displays the imputation of missing data for subject *p* based on the drinking information of subjects with drinking patterns most similar to that of *p*. Similarity in the drinking patterns of two subjects is measured using their *cosine distance*. In this hypothetical example, there are three subjects (*q*, *r* and *s)* for whom estimates of alcohol consumption were collected on three different days during pregnancy. For subject *p* information is missing for the third day. The nearest neighbor for subject p is subject q. The angle between $p'O$ and $q'O$ is zero which means that *p* and *q* have exactly the same drinking pattern, as they both consumed three times more drinks on day 1 than on day 2. The next nearest neighbor for subject p is subject r as the angle between them is small. In practice, it is computationally complex to calculate an angle and we can use the *cosine* as a good approximation. Once the *k* nearest neighbors of *p* are identified, the weighted average of the drinking data of these neighbors for the day for which *p*'s drinking data are missing is taken as the best estimate of the missing data. The weighted average is taken to assure that the neighbors nearer to *p* have more influence on the predicted value than the ones further away from it. We also scaled the imputed values to individual

consumption level. In this example, the scaling adjustment is needed because though *p* and *r* have similar drinking patterns, p is heavier drinker than r. Details of the computation of cosine similarity and scaling adjustment are described in appendix 1.

Data preparation:

We first converted the data to a single record (row) per person, where drinking values were separate variables (columns), one variable for each drinking day starting from day -15 (2 weeks prior to LMP) and ending at day 310 (maximum possible pregnancy length). We used the distance between a fixed date before the start of the study (Saturday, January 1, 2000), and the beginning of pregnancy (i.e., day -15) to find the day of the week the pregnancy started. This was then used to temporally align each subject prior to computation of the cosine distances. For example, when computing the nearest neighbors of a participant *p* whose pregnancy started on a Wednesday, if we encountered another participant *q* whose pregnancy started on a Monday, we aligned day -15 of *p* (a Wednesday) with day -13 of *q* (another Wednesday) and ignored the first two days (days -15 and -14) of *q* and the last two days (days 309 and 310) of *p.* The rationale behind this alignment is that the drinking behavior often varies by the day of the week[15]. We also Winsorized (capped) the outlier drinking values at 3 SD (21 for South Africa and 28 for Northern Plains sites) to reduce the impact of outlier values in determining the imputed values. As the pattern of drinking in subjects with data missing for a large number of days cannot be established; we excluded subjects with more than 200 days missing and those subjects who did not have any data in the first trimester.

Assessment of performance /validation

We validated our approach by comparing actual values from a subset of subjects with no missing data in trimester 1 with imputed values obtained after random deletion of data for 5 to 15 consecutive days. The first trimester was selected for validation because the proportion of women drinking and the magnitude of their drinking is highest in trimester 1, particularly for the days before pregnancy recognition. To evaluate imputation performance, we computed the root mean squared error (RMSE) for the predicted drinking values in the deleted segments as follows,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

where n is the length of a segment and for $1 \leq i \leq n$, y_i and $\hat{y}_i$ are the actual and predicted value, respectively, of the i^{th} entry of the segment. We then calculated overall prediction accuracy of drinking status as proportion of accurate classification and plotted it in a confusion matrix (figure 5, panel b). In addition, we calculated absolute differences between actual and predicted values for drinking and non-drinking days separately.

Results:

Description of missing data

Participants contributed a total of 3.2 million person-days of observation in the study, of which 0.36 million (11.4 %) person-days were missing. Based on the data collected using the TLFB method about 45% of the participants (n=5396) had alcohol use data for every single day of their pregnancy while the remaining 55% (n=6492) had at least 1 day of alcohol-use data missing. Among the study participants 62% (n=7119) were drinkers, i.e., consumed at least 1 drink during pregnancy. Figure 3 shows the distribution of missing days per participant by study site. Overall, Northern Plains sites had fewer missing data, with over 50% of the participants having 30 or fewer days of missing data (figure 3). Most of the missing data in the South Africa site are from the early trimesters which largely reflects later enrollment at that site, while the majority of missing data in the NA site are in the 3rd trimester (data not shown).

Application of k-NN

Length of reference segment:

The largest possible reference segment in the PASS data set is 324 days, the maximum length of the pregnancy (310 days) plus 2 weeks before pregnancy. However, as mentioned in a previous section, women with complete data were more likely to be nondrinkers or light drinkers, hence using them as reference could produce an underestimate of true drinking values. The tradeoff between selecting a larger or smaller segment size is that smaller segment sizes (e.g., 7 days) allows more segments to be included as reference; but the smaller the segment becomes, the less accurate is the algorithm's characterization of specific patterns of drinking. We determined that a reduction of segment sizes below 55 days did not increase available reference segments significantly (Figure 4). Thus, a segment size of approximately 2

months (55 days) retained the majority of the subjects in the reference pool without significantly diminishing the ability to identify their drinking patterns.

Number of neighbors, K

To identify the optimal number of neighbors for imputation, we varied the number of neighbors k from 1 to 10. Figure 5 shows the distribution of root mean square errors (RMSE) as a function of k in both study sites are combined. For the prediction of nondrinking segments, K= 1 provided the lowest RMSE (panel a) and using K>1 (multiple neighbors) provided lower RMSE for the prediction of drinking segments. Although the mean RMSE value in the drinking segments decreased as the value of k is increased, mean RMSE after k =5 did not decrease substantially. Based on their relative performance in classification accuracy (Figure 5: panel b) in both sites, we concluded that k=5 provided reasonable accuracy for both drinking and nondrinking days. We found the k-NN algorithm made exact predictions of drinking status for 76% drinking segments in the combined sample. The algorithm predicted nondrinking status accurately in 74% and 58% of the deleted segments in South Africa and Northern Plains respectively (data not shown). Using K=5, the approach predicted nondrinking segments within +/- 1 drinks for 80.5% of deleted segments in South Africa and 70.6% in the Northern Plains (Figure 5: panel c).

Average drinking after imputation:

Figure 7 shows the mean number of drinks per person by trimester before and after imputation. Following imputation, the mean number of drinks in South Africa increased by an average of 2 drinks in first trimester, while the increase for the Northern Plains sites was just below 1 drink in first trimester. Following imputation, the magnitude of increase in mean drinks in South Africa was higher than that in Northern Plains. The Northern Plains had fewer missing data than the South African site. Consequently, although many individual drinking values were changed, imputation had little effect on the average drinking values in Northern Plains sites.

Discussion:

The objective of this report was to describe the application of a machine learning algorithm to impute missing daily alcohol consumption data in a prospective study among pregnant women. When pregnant women were asked about recent alcohol consumption during their prenatal visits, days of missing data were an inherent consequence of the assessment methodology; and, there were more missing data among recent drinkers. Thus, missing data points were not at random. We implemented an extension of a kNN algorithm which accounted for the absence of a 'typical/classic' reference group, i.e., training data present report is the first to describe this method to impute

missing alcohol consumption data in a longitudinal study among pregnant women. Validation of our approach showed high agreement between actual and predicted drinking values.

There is a paucity of studies on missing data techniques and their statistical validity in alcohol and drug use research studies [16]. Published work has not yet reported the performance of any machine learning method for imputation of missing alcohol data. In a simulated dataset, Hallgren et al. compared methods of imputation including complete case analysis, last observation carried forward, the worst-case scenario of missing equals any drinking or heavy drinking, multiple imputation (MI), full information maximum likelihood (FIML) and concluded that MI and FIML yielded less biased estimates [17, 18]. A recent study by Grittner et al. also found MI produced least bias based on their work in a longitudinal study in Denmark with five alcohol measurements over a period of five years [19]. However, all methods in the study including the MI produced an underestimate of the actual drinking level. In addition, MI models are originally recommended for imputation of a single value per subject [20]. To impute irregularly spaced missing longitudinal data as in PASS, complex extensions of MI would be needed [21].

There are several advantages with using a non-parametric algorithm such as the kNN algorithm for imputation of missing data. The majority of standard software packages rely on the assumption of normal distribution of multivariate data, therefore imputation of repeated longitudinal data in most software options is challenging [21]. In PASS, alcohol data were collected at the daily level resulting in a high total volume of both data per participant and missing data. In the general population, alcohol consumption in pregnancy is highly skewed with the majority of the drinking concentrated in the first trimester. We observed this pattern in PASS, however, there was also a gradually decreasing drinking pattern among many study subjects. In such scenarios, a nonparametric method such as kNN has the advantage of not making a distributional assumption.

The kNN algorithm is increasingly used to impute missing data in research with high volume data such as genetics and metabolomics studies [22, 23]. In several recent reports the kNN algorithm was shown to produce the smallest imputation error compared to methods such as mean and median imputation, Bayesian linear regression, K-Means, K-Medoids clustering algorithms [24, 25]. However, some studies reported that simpler methods such as mean or median replacement were as adequate as methods like kNN when imputation was followed by clustering of genetic data [26]. On the other hand, some have reported slightly better performance of random forest over kNN to impute metabolomics data [27]. Another study noted improvement of performance of kNN when additional information such as SES and demographic data were included in the prediction model [28]. The validity and accuracy of imputation methods will likely vary with the data type, data structure, mechanism of missingness and amount of missing data. Therefore, future studies need to evaluate the performance of different machine learning algorithms to impute alcohol consumption data.

Conclusions:

The *K*-NN method can be used to impute missing prospectively collected prenatal drinking data with high accuracy. The performance of this algorithm may vary in other data sets depending on the amount missing data and amount of drinking in the population. The *K*-NN algorithm can be used to impute missing similar data such as smoking and substance abuse data in longitudinal studies.

Abbreviations

K-NN
K Nearest Neighbor
LMP
Last menstrual period
MI
multiple imputation
PAE
Prenatal alcohol exposure
PASS
Prenatal Alcohol and SIDS and Stillbirth Network
SES
Socioeconomic status
SIDS
sudden infant death syndrome
TLFB
Timeline Followback Method
FIML
full information maximum likelihood
RMSE
Root mean square errors

Declarations

Ethics approval and consent to participate

Ethical approval was obtained from Stellenbosch University, Sanford Health, the Indian Health Service and from participating Tribal Nations.

Consent for Publication

Written informed consent was obtained from all participants.

The datasets used and during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by grants UH3OD023279, U01HD055154, U01HD045935, U01HD055155, and U01AA016501, issued by the Office of the Director, National Institutes of Health of the National Institutes of Health, National Institute on Alcohol Abuse and Alcoholism, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, and the National Institute on Deafness and Other Communication Disorders. The opinions expressed in this paper are those of the authors and do not necessarily represent the official views of the National Institutes of Health, the Eunice Kennedy Shriver National Institute of Child Health and Development, the National Institute on Alcohol Abuse and Alcoholism, or the National Institute on Deafness and Other Communication Disorders.

Author contribution

Ayesha Sania conceptualized and conducted the data analysis, interpreted the results and wrote the first draft of the manuscript. Nicolò Pini, Michael M. Myers participated in the data analyses and contributed in the manuscript writing. Lauren C. Shuffrey and Maristella Lucchini contributed in interpretation of the results and manuscript writing. Hein J. Odendaal and Morgan E. Nelson participated in study implementation and data collection and provided critical inputs on the manuscript. Amy J. Elliott and William P. Fifer are the principal investigators of the Safe Passage Study and contributed to the study design, implementation, analysis and interpretation of the data. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

1. Dawson DA. Methodological issues in measuring alcohol use. Alcohol Res Health. 2003;27(1):18–29.

2. Feunekes GI, van 't Veer P, van Staveren WA, Kok FJ. Alcohol intake assessment: the sober facts. Am J Epidemiol. 1999;150(1):105–12.
3. Buu A, Yang S, Li R, Zimmerman MA, Cunningham RM, Walton MA. Examining measurement reactivity in daily diary data on substance use: Results from a randomized experiment. Addict Behav. 2020;102:106198.
4. McQuire C, Paranjothy S, Hurt L, Mann M, Farewell D, Kemp A. **Objective Measures of Prenatal Alcohol Exposure: A Systematic Review**. *Pediatrics* 2016, 138(3).
5. O'Keeffe LM, Kearney PM, McCarthy FP, Khashan AS, Greene RA, North RA, Poston L, McCowan LM, Baker PN, Dekker GA, et al. Prevalence and predictors of alcohol use during pregnancy: findings from international multicentre cohort studies. BMJ Open. 2015;5(7):e006323.
6. Rubin D. Inference and missing data. Biometrika. 1976;63(3):581–92.(581–592).
7. Simkhada B, Teijlingen ER, Porter M, Simkhada P. Factors affecting the utilization of antenatal care in developing countries: systematic review of the literature. J Adv Nurs. 2008;61(3):244–60.
8. Skagerstrom J, Chang G, Nilsen P. Predictors of drinking during pregnancy: a systematic review. J Womens Health (Larchmt). 2011;20(6):901–13.
9. Dukes KA, Burd L, Elliott AJ, Fifer WP, Folkerth RD, Hankins GD, Hereld D, Hoffman HJ, Myers MM, Odendaal HJ, et al. The safe passage study: design, methods, recruitment, and follow-up approach. Paediatr Perinat Epidemiol. 2014;28(5):455–65.
10. Dukes K, Tripp T, Petersen J, Robinson F, Odendaal H, Elliott A, Willinger M, Hereld D, Raffo C, Kinney HC, et al. A modified Timeline Followback assessment to capture alcohol exposure in pregnant women: Application in the Safe Passage Study. Alcohol. 2017;62:17–27.
11. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theor. 2006;13(1):21–7.
12. Elliott P, Hawthorne G. Imputing missing repeated measures data: how should we proceed? Aust N Z J Psychiatry. 2005;39(7):575–82.
13. Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, Marrero J, Zhu J, Higgins PD. **Comparison of imputation methods for missing laboratory data in medicine**. *BMJ Open* 2013, 3(8).
14. Brick J. Standardization of alcohol calculations in research. Alcohol Clin Exp Res. 2006;30(8):1276–87.
15. Room R, Makela P, Benegal V, Greenfield TK, Hettige S, Tumwesigye NM, Wilsnack R. Times to drink: cross-cultural variations in drinking in the rhythm of the week. Int J Public Health. 2012;57(1):107–17.
16. Grigsby TJ, McLawhorn J. Missing Data Techniques and the Statistical Conclusion Validity of Survey-Based Alcohol and Drug Use Research Studies: A Review and Comment on Reproducibility. Journal of Drug Issues. 2018;49(1):44–56.
17. Hallgren KA, Witkiewitz K. Missing data in alcohol clinical trials: a comparison of methods. Alcohol Clin Exp Res. 2013;37(12):2152–60.

18. Hallgren KA, Witkiewitz K, Kranzler HR, Falk DE, Litten RZ, O'Malley SS, Anton RF. In conjunction with the Alcohol Clinical Trials Initiative W: **Missing Data in Alcohol Clinical Trials with Binary Outcomes**. Alcohol Clin Exp Res. 2016;40(7):1548–57.
19. Grittner U, Gmel G, Ripatti S, Bloomfield K, Wicki M. Missing value imputation in longitudinal measures of alcohol consumption. Int J Methods Psychiatr Res. 2011;20(1):50–61.
20. Rubin D. Multiple imputation for Nonresponse in surveys. New York: Wiley; 1987.
21. Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Med Res Methodol. 2018;18(1):168.
22. Liao SG, Lin Y, Kang DD, Chandra D, Bon J, Kaminski N, Sciurba FC, Tseng GC. Missing value imputation in high-dimensional phenomic data: imputable or not, and how? BMC Bioinformatics. 2014;15:346.
23. Shah JS, Rai SN, DeFilippis AP, Hill BG, Bhatnagar A, Brock GN. Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. BMC Bioinformatics. 2017;18(1):114.
24. Jadhav A, Pramod D, Ramanathan K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence. 2019;33(10):913–33.
25. Mahboob T, Ijaz A, Shahzad A, Kalsoom M: **Handling Missing Values in Chronic Kidney Disease Datasets Using KNN, K-Means and K-Medoids Algorithms**. In: 2018 *12th International Conference on Open Source Systems and Technologies (ICOSST): 19–21 Dec. 2018 2018*; 2018: 76–81.
26. de Souto MC, Jaskowiak PA, Costa IG. Impact of missing data imputation methods on gene expression clustering and classification. BMC Bioinformatics. 2015;16:64.
27. Kokla M, Virtanen J, Kolehmainen M, Paananen J, Hanhineva K. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. BMC Bioinformatics. 2019;20(1):492.
28. Schwender H. Imputing Missing Genotypes with Weighted k Nearest Neighbors. Journal of Toxicology Environmental Health Part A. 2012;75(8–10):438–46.

Figures

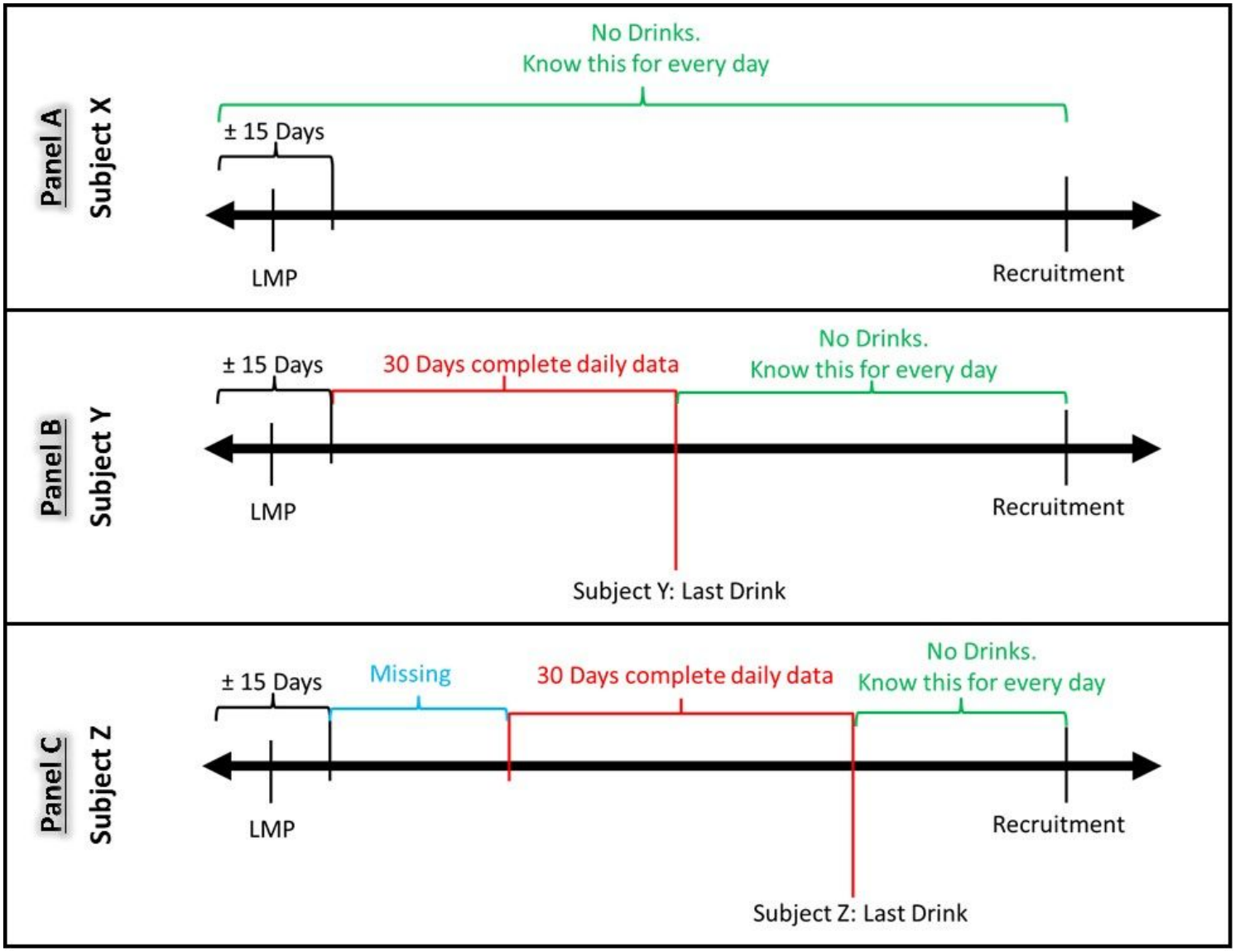

Figure 1

Timing of alcohol consumption during pregnancy and its relation to missing data Let's assume that subjects X, Y and Z were enrolled at the same gestational ages for their respective pregnancies. The alcohol consumption of Subject X is depicted in Panel A. Participant X is a non-drinker given no alcohol consumption is reported in the time interval which spans from LMP and recrutiment. Both Subject Y (Panel B) and Subjects Z (Panel C) did report at least an event of alcohol consumption in the same interval. Neverthless, the timing of alcohol intake is different for the participants, thus resulting is the absence (Subject Y) and presence (Subject Z) of data missingess. Considering Subject Y, the time interval between last alcohol intake and LMP is less or equal 30 days, thus there is no gap in alcohol consumption information, resulting in a complete timeline from recruitment back to LMP. On the contrary, Subject Z reported her last drinking event more recently with respect to Subject Y, thus the interval between last alcohol consumption and LMP is greater than 30 days. In this latter case, we have data

ment.

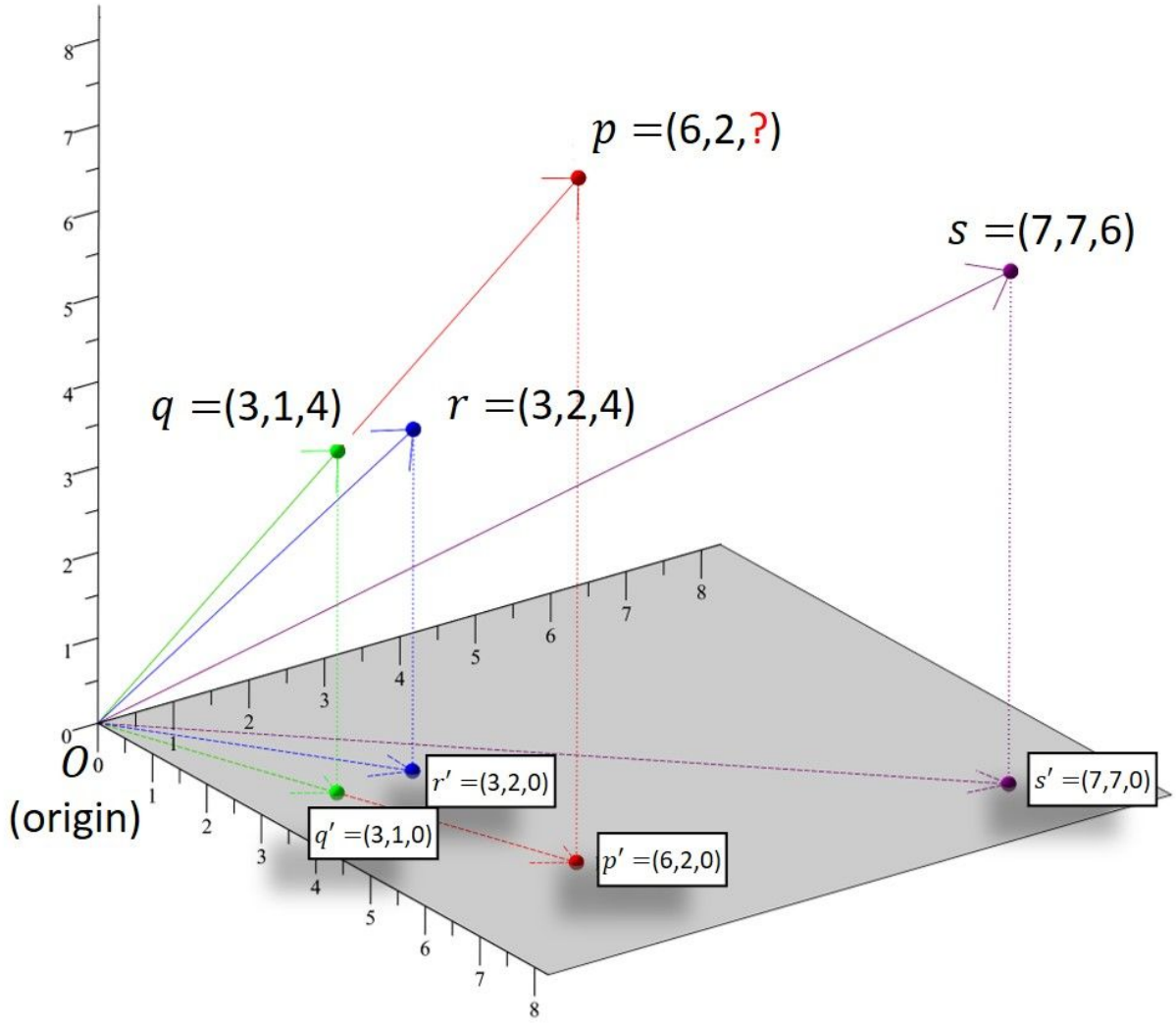

Figure 2

Subjects p, q, r and s are mapped to points p', q', r' and s', respectively, in 2-dimensional space based on the two days for which data are available for all of them. If a subject had x drinks on day one and y drinks on day two then it is mapped to point (x, y, 0) on the 2-dimensional xy plane.

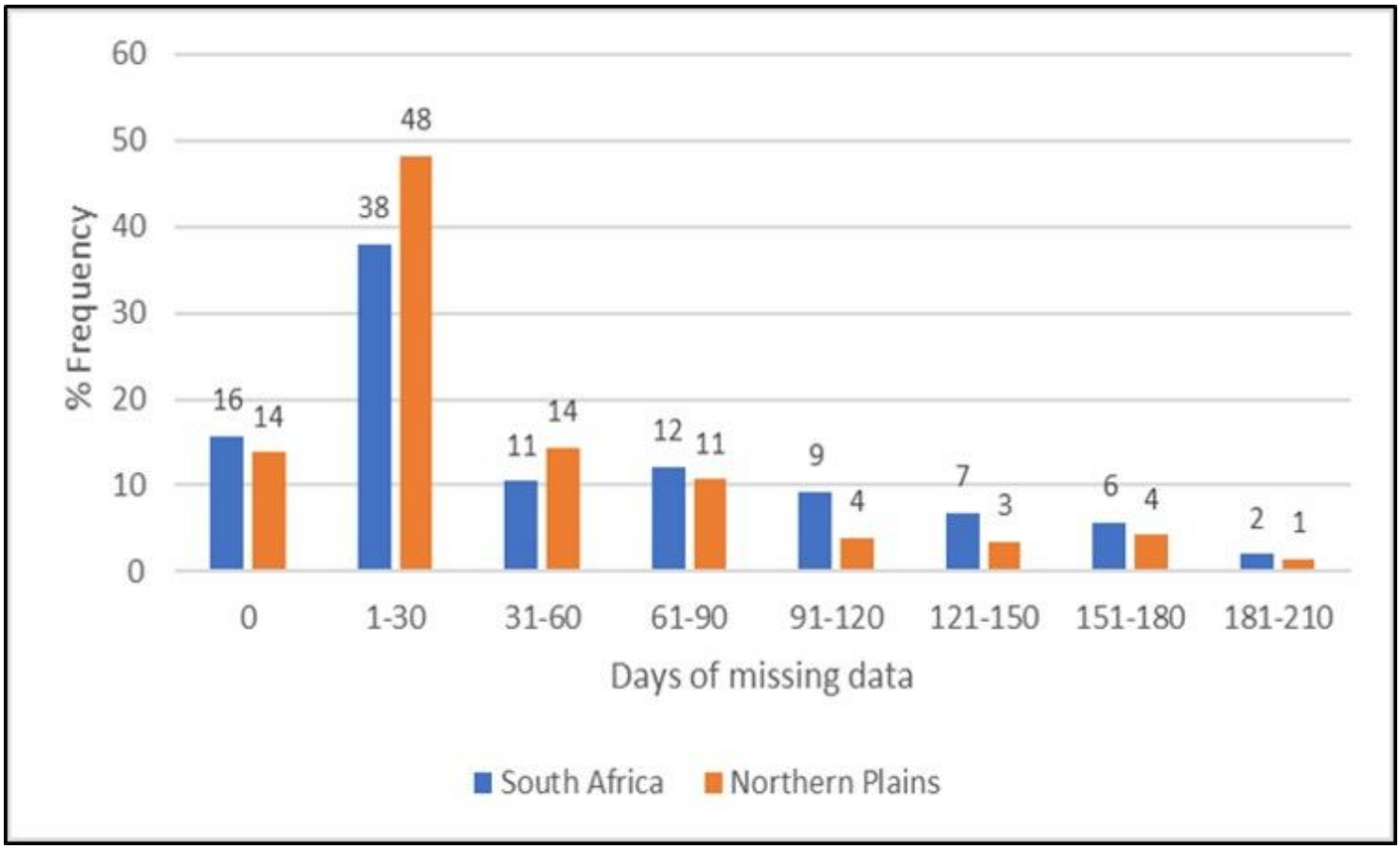

Figure 3

Distribution of missing data by study site

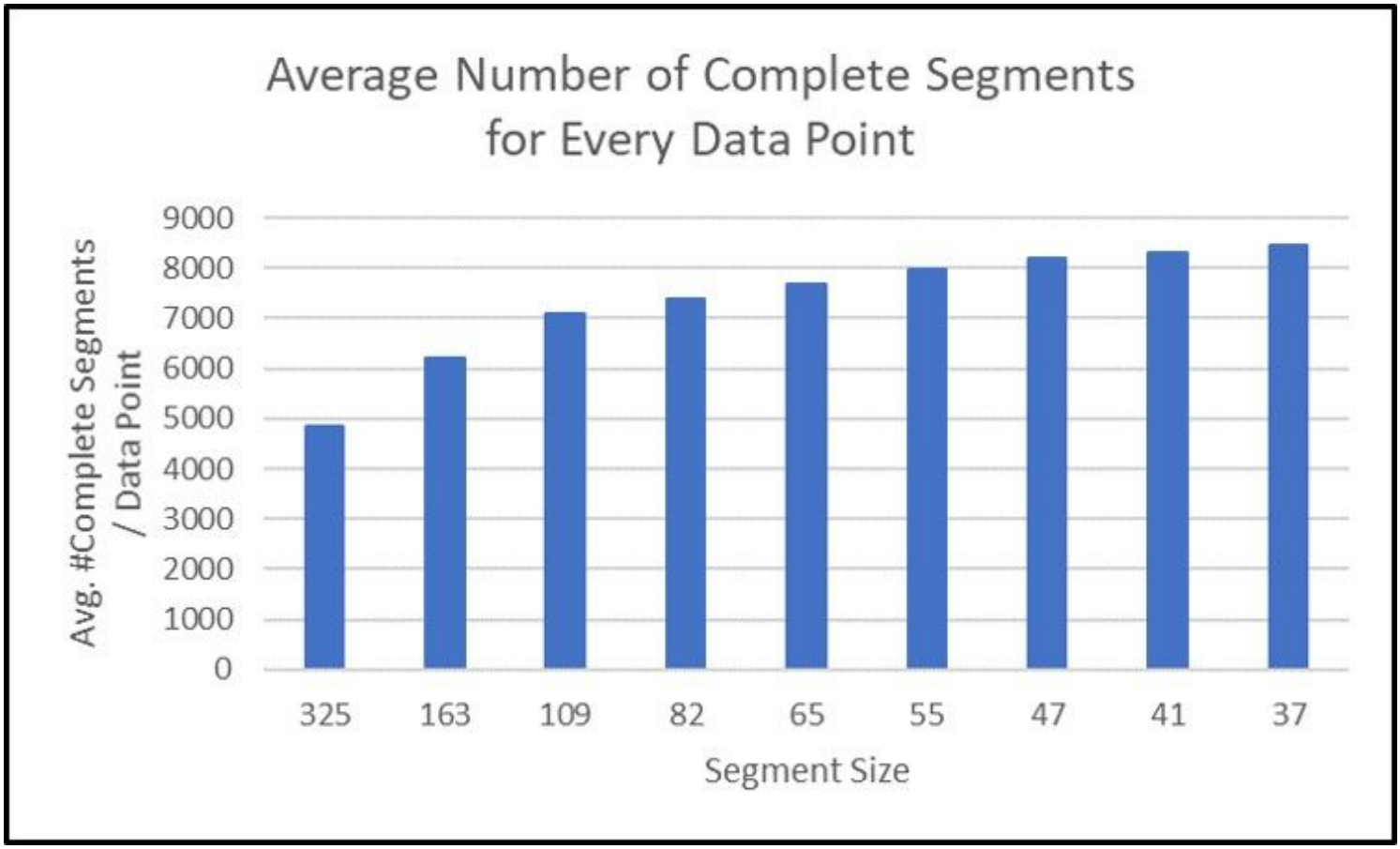

Figure 4

Average number of complete segments for each segment length

Figure 5(a)

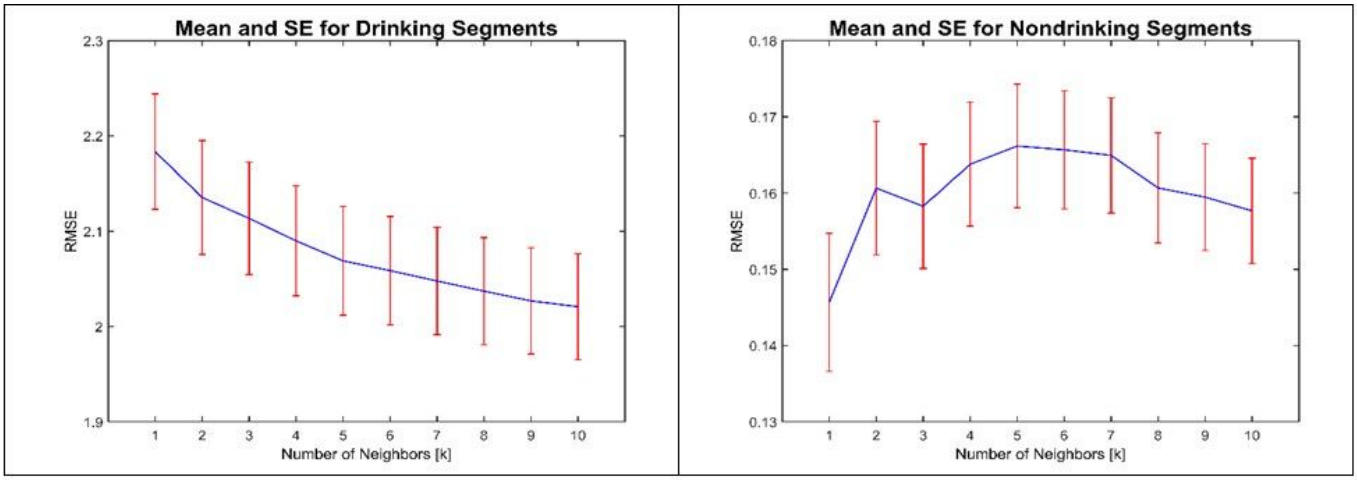

Figure 5(b)

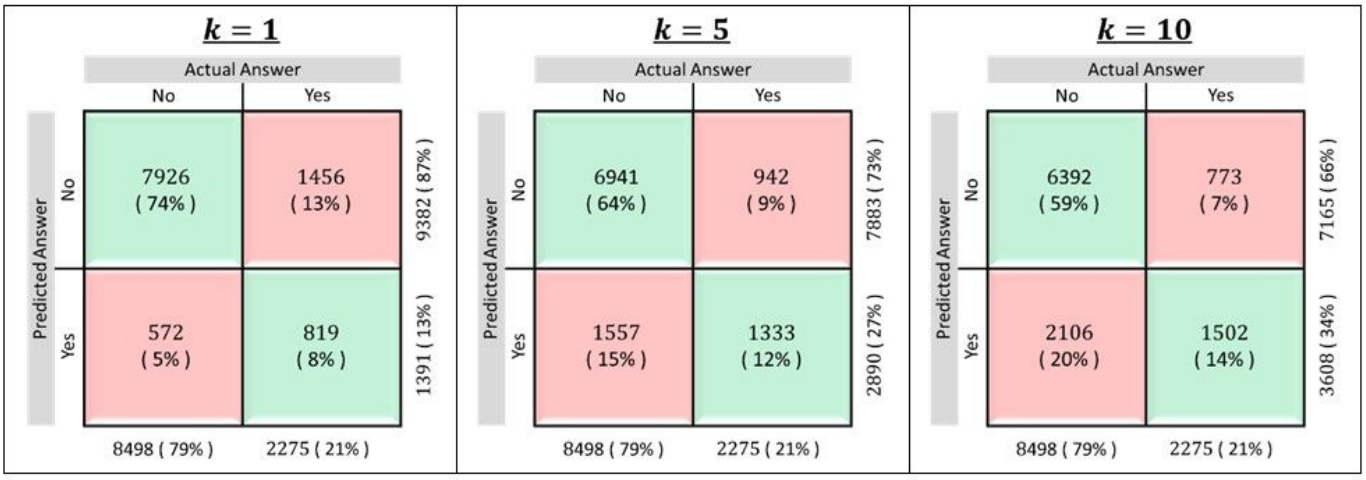

Figure 5(c)

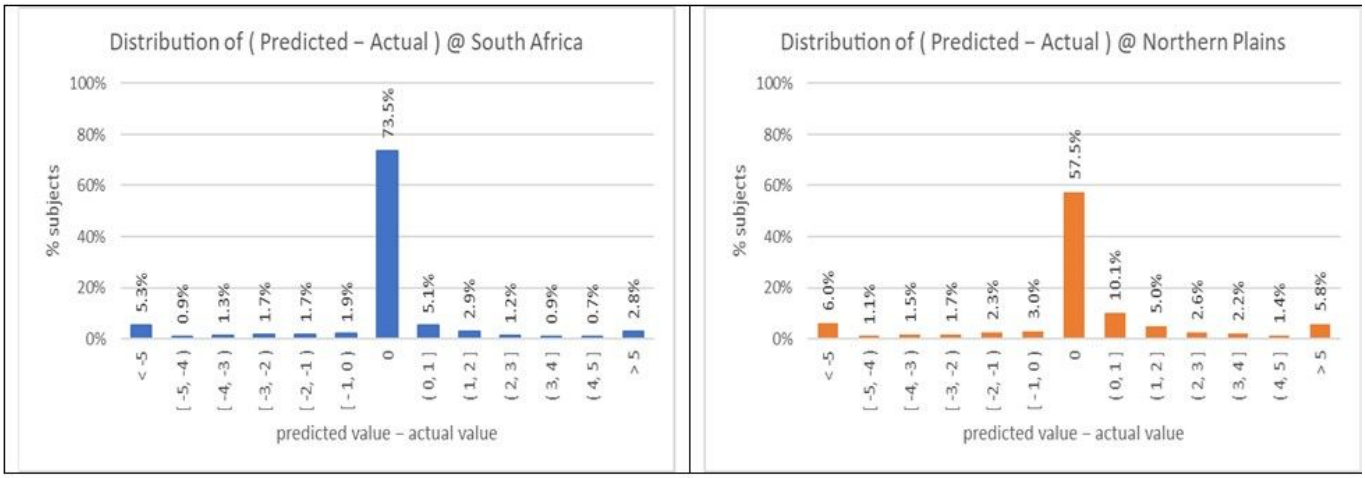

Figure 5

Figure 5, panel a: Distribution of RMSE in drinking and nondrinking segments Figure 5, panel b: Classification accuracy for k=1, 5, and 10 Figure 5, panel c: Difference between predicted and actual values by study site

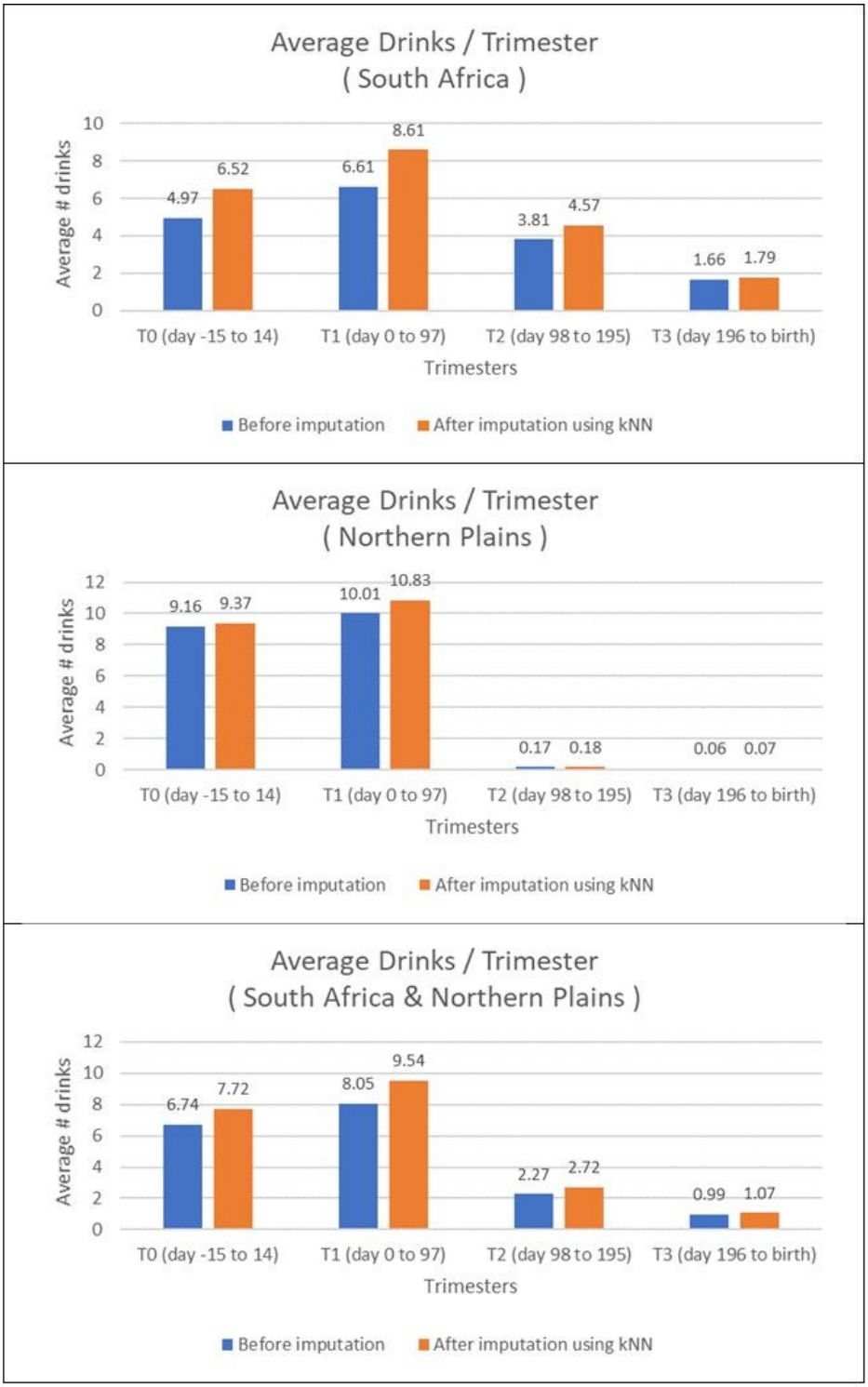

Figure 6

Average drinks per trimester before and after imputation

Supplementary Files

Loading [MathJax]/jax/output/CommonHTML/jax.js ed with this preprint. Click to download.

- Appendix1.docx