UNC
ADAMS SCHOOL OF DENTISTRY

Development of a Novel Imputation Method for Missing Fluoride Measurements in a Community-Based Epidemiologic Study

Gu Y[1], Shah M[2], Shrestha P[2,3], Simancas-Pallares M[2], Karhade DS[2], Ginnis J[2], Divaris K[2,3]

[1] *Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC*
[2] *Division of Pediatric and Public Health, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC*
[3] *Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC*

Abstract

Objectives: Direct measurements of domestic water fluoride content provide valuable information regarding individuals' fluoride exposure; however, they are rare and logistically challenging to carry out at a large scale. Here we present the development and evaluation of a novel method for the imputation of missing domestic water fluoride concentration data informed by spatial autocorrelation.

Methods: We used domestic water fluoride concentration data that were generated in ZOE 2.0, an epidemiologic study of early childhood oral health in NC. Fluoride concentration was measured with the EPA 300.0 method in domestic water samples that were available for approximately 25% of study participants. Residential locations were geocoded using ArcGIS Pro 2.2 software. Additional information used included questionnaire responses on home water source and clinical data on children's dental caries status [early childhood caries (ECC) case status]. We initially considered 3 existing interpolation methods including inverse distance weighting (IDW), universal kriging (UK), and k-nearest neighbors (KNN). The new method (PAMRF) was based on a combination of partitioning around medoids (PAM) clustering and Random Forest classification.

Results: A leave-one-out cross-validation (LOOCV) suggested that, based on error rates, PAMRF outperforms the other 3 methods. One-third of the 4,779 missing fluoride values were imputed with PAMRF with ⩾95% confidence. The estimate of association between optimal fluoride concentration (⩾0.60 ppm) and ECC remained virtually unchanged between analyses of observed and observed plus all imputed data, but there were large gains in precision: observed (n=1,360)—prevalence ratio (PR)=0.86 [95% confidence interval (CI)=0.77–0.95], P=2×10^{-3} vs. observed and all imputed (n=5,761)—PR=0.86 (95% CI=0.83–0.91), P=1×10^{-10}.

Conclusion: We have developed a powerful method for imputing missing fluoride values that outperforms existing approaches. Investigators can use PAMRF to select different sets of imputed values according to tuning parameters, allowable error rates or target sample size, depending on the requirements of their application.

Funding: NIH/NIDCR U01-DE025046

Introduction

Despite the valuable information they can provide, direct measurements of domestic water fluoride content are challenging to carry out at a large scale. Here, we used domestic water fluoride concentration data that were generated in ZOE 2.0, a large-scale community-based epidemiologic study of early childhood oral health in North Carolina (Ginnis et al., 2019), wherein ~75% of study participants had missing fluoride concentration information.

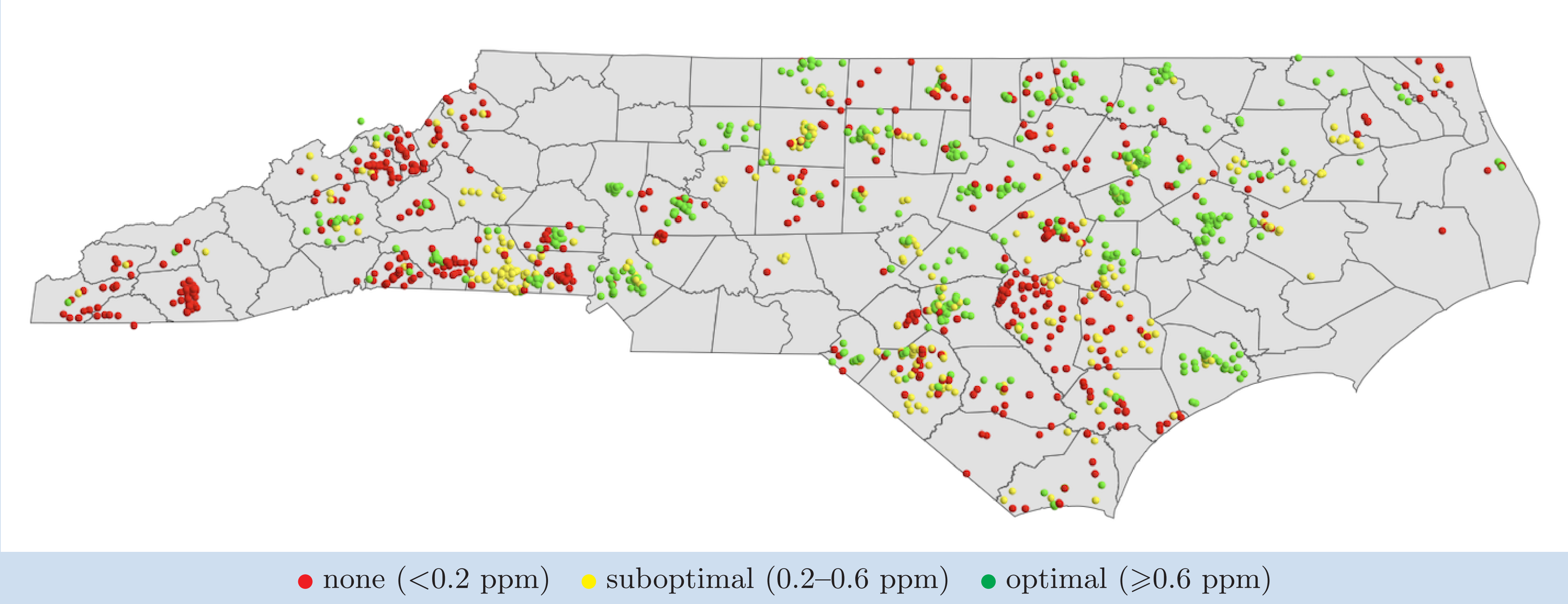

Figure 1. *Geographical distribution of fluoride concentration levels, points representing the geocoded residential locations for participants in the study, missing values not included.*

Given the strong spatial autocorrelation of fluoride concentration levels, our motivation was to impute the missing values based on the measured values from families' closest neighbors. We also used additional information from questionnaire responses on families' home water source, since it strongly correlated with the measured fluoride concentration.

Results & Discussion

Table 1. *LOOCV error rate of each method, with optimal values selected for all tuning parameters using LOOCV.*

Method	KNN	IDW	UK	PAMRF
Error rate	0.2372	0.3411	0.3611	0.2102

Table 2. *Correlations between predicted results using different imputation methods based on LOOCV and true values.*

	True value	IDW	UK	KNN	PAMRF
True value	1.0000	0.5049	0.6095	0.6690	0.7127
IDW	0.5049	1.0000	0.6457	0.4171	0.5974
UK	0.6095	0.6457	1.0000	0.5323	0.6547
KNN	0.6690	0.4171	0.5323	1.0000	0.7331
PAMRF	0.7127	0.5974	0.6547	0.7331	1.0000

Based on the results of leave-one-out cross-validation (LOOCV), PAMRF outperforms the other three methods, with the lowest error rate and the highest correlation with true values. Nearly one-third of the 4,779 missing fluoride values were imputed with PAMRF with ⩾95% confidence. By setting an appropriate threshold of minimum vote, one can achieve error rates and target sample sizes that are specifically allowable in each application problem. The estimate of association between optimal fluoride concentration (⩾0.60 ppm) and ECC remained virtually unchanged between analyses of observed and observed plus all imputed data, but there were large gains in precision: observed (n=1,360)—prevalence ratio (PR)=0.86 [95% confidence interval (CI)=0.77–0.95], P=2×10^{-3} vs. observed and all imputed (n=5,761)—PR=0.86 (95% CI=0.83–0.91), P=1×10^{-10}.

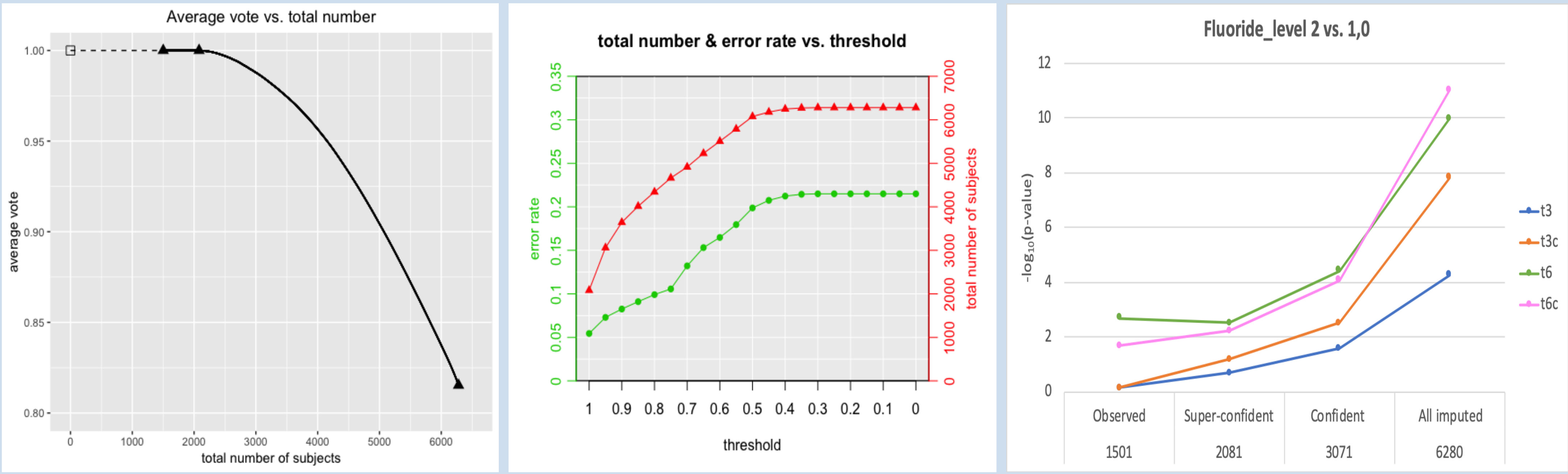

Figure 2. *The first plot displays the function of average vote among imputed values versus total sample size. The second plot shows patterns of LOOCV error rate and total sample size over different thresholds. The third plot presents the significance of association between optimal fluoride concentration and ECC, based on four different sets of data.*

References

J. Ginnis, A. G. F. Zandoná, G. D. Slade, J. Cantrell, M. E. Antonio, B. T. Pahel, B. D. Meyer, P. Shrestha, M. A. Simancas-Pallares, A. R. Joshi, et al. Measurement of early childhood oral health for research purposes: Dental caries experience and developmental defects of the enamel in the primary dentition. In *Odontogenesis*, pages 511–523. Springer, 2019.

Methods

Let $Z(\cdot)$ denote the continuous fluoride concentration value, $\boldsymbol{s}_0$ denote the interpolated point, and $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ denote the n closest neighbors of $\boldsymbol{s}_0$. We considered the following three existing interpolation methods as well as our new imputation method.

1. Inverse Distance Weighting (IDW). The missing fluoride value can be imputed by

$$\widehat{Z}(\boldsymbol{s}_0) = \frac{\sum_{i=1}^{n} Z(\boldsymbol{s}_i)/d(\boldsymbol{s}_0, \boldsymbol{s}_i)^p}{\sum_{i=1}^{n} 1/d(\boldsymbol{s}_0, \boldsymbol{s}_i)^p} \tag{1}$$

where $d(\cdot,\cdot)$ is the geographical distance between two points, p is the power parameter.

2. Universal Kriging (UK). We impute the missing fluoride value by

$$\widehat{Z}(\boldsymbol{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{s}_i) \tag{2}$$

Under the gaussian process regression model and the assumption of distance-dependent covariogram $C(h)$, the optimal weights λ_i's are given by

$$\left\{\boldsymbol{c} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x} - \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{c})\right\}'\boldsymbol{\Sigma}^{-1} \tag{3}$$

where $\boldsymbol{c} = (C(\boldsymbol{s}_0, \boldsymbol{s}_1), \ldots, C(\boldsymbol{s}_0, \boldsymbol{s}_n))'$, $\boldsymbol{\Sigma}$ is an $n \times n$ matrix whose (i,j)-th element is $C(\boldsymbol{s}_i, \boldsymbol{s}_j)$, $\boldsymbol{X}$ is the matrix of covariates for $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, and $\boldsymbol{x}$ is the vector of covariates for $\boldsymbol{s}_0$. The estimated covariogram can be obtained by modeling the semivariogram with various models.

3. K-Nearest Neighbors (KNN). We impute the missing fluoride value by the majority vote of the k closest neighbors' fluoride levels. Here we calculate the Euclidean distance between two subjects based on their X and Y coordinates of geocoded residential locations, as well as their home water sources.

4. PAMRF. Our new method is based on a combination of partitioning around medoids (PAM) clustering and Random Forest classification. Within each county, we first divide all the subjects into different clusters using PAM clustering, based on their fluoride levels, X and Y coordinates, and home water sources. Then we train a random forest model which accounts for the effects of residential location and water source on the cluster to which a subject belongs. We assign each subject with a missing value into a specific cluster and impute the missing value by the fluoride level of that cluster's medoid.

Conclusion

We have developed a powerful method for imputing missing fluoride values that outperforms existing approaches. While the estimate of association between optimal fluoride concentration and ECC remain virtually unchanged, we noted large gains in precision and statistical significance when the imputed data were included in analyses. Investigators can use PAMRF to select different sets of imputed values by setting different thresholds, depending on the requirements regarding error rate and sample size in their applications. In future studies, we may consider other measures of distance such as Manhattan and Gower, as well as study the influence of the distance function on the imputation performance of PAMRF. We may further explore how to incorporate multiple imputation into PAMRF to avoid underrepresentation of uncertainty in a single imputation process.

Acknowledgements

Research reported in this presentation was supported by the National Institute of Dental & Craniofacial Research of the National Institute of Health (NIH) under award number U01DE025046. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.