IMPROVING IMPUTATION ACCURACY

U.S. Census Bureau | Public Sector Statistical Methods Branch

Peter Kirgis, Data Science Fellow | *Princeton Univeristy, Master of Public Affairs*

Keywords:

data science, imputation, estimation, regression, research

Summary:

As a Coding it Forward Fellow, Peter conducted research to evaluate and explore alternatives to an existing methodology used to impute data for non-respondent local governments in a Census survey. As a first step in this work, Peter created maps, histograms, and scatterplots to visualize key metrics for the imputation program. Next, Peter designed a **sorting algorithm** to optimize a population parameter that groups local governments into imputation cells. Finally, Peter designed, tested, and visualized the results of both **parametric and non-parametric regression** approaches to directly estimate all missing data in a given survey year. This research serves as the jumping-off point for a long-term imputation modernization effort at the Public Sector Statistical Methods Branch.

IMPROVING IMPUTATION ACCURACY

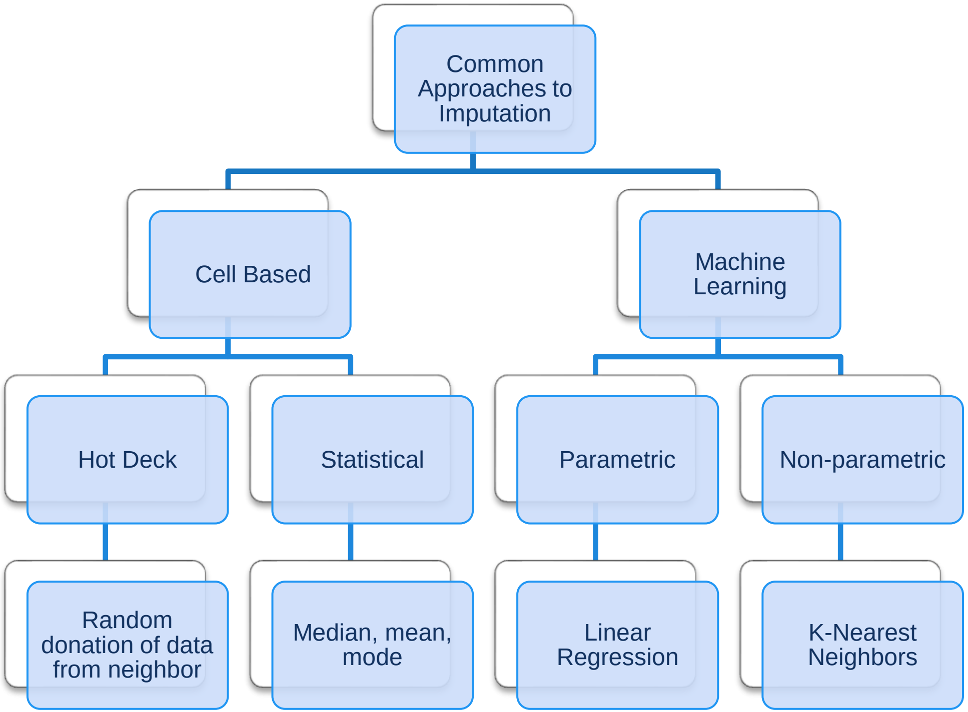
Public Sector Statistical Methods Branch
U.S. Census Bureau
Erica Marquette — Branch Chief

PETER KIRGIS
Princeton University
Master in Public Affairs

The Annual Survey of Local Finances (ALFIN) currently relies on a **geographic, cell-based, statistical** strategy for imputing missing data.

cif>

BACKGROUND



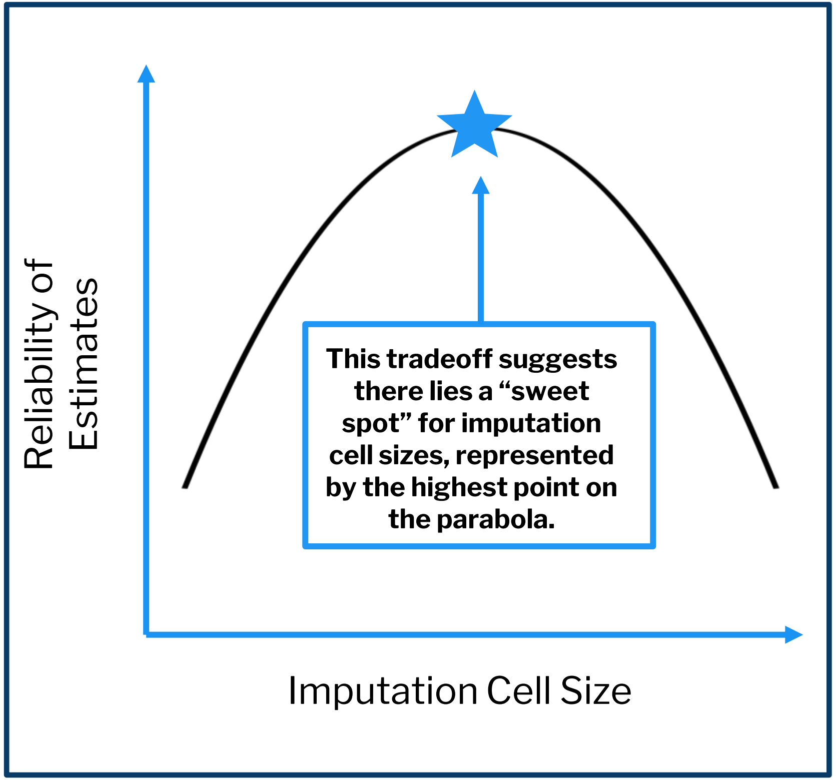
Example ALFIN Imputation Cells

State	Gov't Type	Population Band
Alabama	County	ALL
Alabama	Municipality	Population ≥ 20k
Alabama	Municipality	20k > Population ≥ 5k
Alabama	Municipality	5k > Population ≥ 1k
Alabama	Municipality	Population < 1k

Note: sample categories not taken from database

PROBLEM DEFINITION

- Cell-based imputation methods face a tradeoff between **homogeneity** and **sample size**.
- With small cell sizes, our statistics (e.g. mean or median) will be sensitive to year-over-year anomalies in the data.
- With large cell sizes, we risk grouping heterogenous units which are unlikely to share financial characteristics.



Can data science techniques help us **incrementally improve or overhaul** this methodology to produce more accurate imputations?

METHODS

1. **Cell Optimization Approaches**
   - Manual: determine the lowest performing cells and test incremental adjustments
   - Algorithmic: generate population bands that most evenly distribute respondent units within a single state and government type for a given cell size floor
2. **Machine Learning Approaches**
   - Linear Regression: use unit information and financial data to predict variables
   - K-Nearest Neighbors: find a small number of units that minimize the distance to our target unit and take the mean of those units weighted inversely by their distance from the target

CELL OPTIMIZATION EVALUATION

Framework: If we can increase cell sizes and/or cell response rates without decreasing cell accuracy, then we have made an improvement to the imputation cell methodology.

Criteria

1. Cell Size – how evenly distributed are units with the cells?
2. Response Rates – are there cells with particularly high rates of non-responding units?
3. Accuracy/Loss – what is the variability of values for respondent units relative to the cell median?

Mean Absolute Error (MAE) = $\frac{1}{n}\sum_{i=1}^{n} |respondent\ value_i - cell\ median\ value|$

CELL OPTIMIZATION FINDINGS

Both the manual and algorithmic approaches indicated room for improvement to the cell parameter used in the existing imputation methodology.

1. **Manual Approach**
   - Eliminated our lowest performing cells measured by response rate and cell size
   - Generated slight decreases in accuracy measured by mean absolute error
2. **Algorithmic Approach**
   - Generated a significant, positive shift in response rates and cell sizes
   - Did not come at the cost of accuracy measured by mean absolute error

MACHINE LEARNING FINDINGS

K-nearest neighbors (KNN) regression more consistently estimated item values across the scale of the data than linear regression.

- K-nearest neighbors and linear regression performed similarly as measured by R^2 and MSE.
- KNN was superior in estimating small item values.
- Both regression approaches were limited by the **sparsity** and **variance** of input data.
- Future analysis is necessary to directly compare the results of direct variable imputation to cell-based, growth rate imputation.

NEXT STEPS

The conclusions from this presentation would be strengthened by the inclusion of nonresponse bias techniques, hybrid modeling approaches, and coding of business rules.

1. **Algorithmic Optimization**
   - Test different variations of the sample size floor
   - Try different sample sizes for different government types
2. **K-Nearest Neighbors Regression**
   - Pre-processing: Account for nonresponse bias using propensity scores
   - Post-processing: Use business logic to clip predictions by state-level differences
   - Build a hybrid model that combines logistic regression and KNN regression

THANK YOU!

Special thanks to my supervisor, Erica Marquette, for her support and assistance!