Health & Place 91 (2025) 103411

Contents lists available at ScienceDirect

Health and Place

journal homepage: www.elsevier.com/locate/healthplace

A multi-constraint Monte Carlo Simulation approach to downscaling cancer data

Lingbo Liu[a], Lauren Cowan[b], Fahui Wang[c,*], Tracy Onega[b,**]

[a] Center for Geographic Analysis, Harvard University, MA, 02138, USA
[b] Department of Population Health Sciences, University of Utah, Huntsman Cancer Institute, Salt Lake City, UT, 84112, USA
[c] Department of Geography and Anthropology, Louisiana State University, LA, 70803, USA

ARTICLE INFO

Keywords:
Cancer data suppression
Spatial data interpolation
Geo-imputation
Population downscaling
Multi-constraint Monte Carlo simulation method

ABSTRACT

This study employs an innovative multi-constraint Monte Carlo simulation method to estimate suppressed county-level cancer counts for population subgroups and extend the downscaling from county to ZIP Code Tabulation Areas (ZCTA) in the U.S. Given the known cancer counts at a higher geographic level and larger demographic groups at the same geographic level as constraints, this method uses the population structure as probability in the Monte Carlo simulation process to estimate suppressed data entries. It not only ensures consistency across various data levels but also accounts for demographic structure that drives varying cancer risks. The 2016–2020 cancer incidence data from the Utah Cancer Registry is used to validate our approach. The method yields results with high precision and consistency across the full urban-rural continuum, and significantly outperforms several machine-learning models such as Random Forest and Extreme Gradient Boosting.

1. Introduction

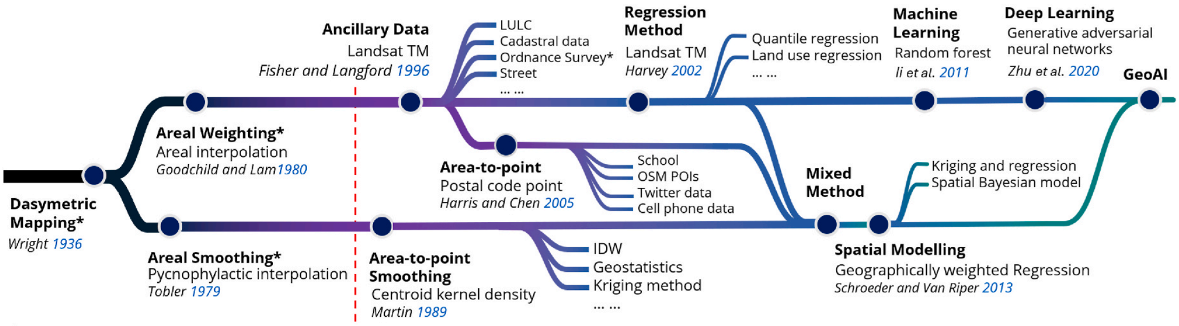
Analyzing disease data at appropriate geographical scales is crucial for understanding spatial patterns of disease burden and informing relevant public policy (Taparra et al., 2022). However, limitations in data collection and privacy protection requirements often result in data aggregated at coarser scales and with missing values. Such an issue is particularly prevalent in cancer datasets (Cook et al., 2021; Amitha et al., 2021), yet estimates at smaller geographic units are often needed for population-level measurement and analyses. Therefore, developing reliable methods to estimate those suppressed values is valuable for spatial analysis at sharper resolutions (Kim et al., 2024) and enables precision public health interventions (Naumova, 2022).

There have been significant advancements in interpolating missing data (Wang et al., 2020) and downscaling population (Wan et al., 2023) over the past decades. However, health (e.g., cancer) data downscaling faces unique challenges (Sahar et al., 2019). Cancer data suppression is geographically nested, and the level of suppression correlates with geographic precision (Buchin et al., 2012). While data at higher levels might be reliable, the proportion of suppressed data increases with finer population groupings and geographic scales. Traditional regression methods for handling missing data often fail to satisfy total volume constraints at higher geographic levels (Howlader et al., 2012). Population downscaling techniques, such as areal interpolation or dasymetric mapping, need to account for the relationship between auxiliary data and cancer data (Walter et al., 2013). Incorrect auxiliary data can lead to erroneous results, e.g., overestimated cases in urban areas and underestimated cases in rural areas (Bozigar et al., 2020). Methods based on Monte Carlo (MC) simulations show promise by estimating cancer incidences proportionally to specific population groups (Shi et al., 2013; Luo et al., 2010). However, more work is needed to account for the nested structure in area units and risk factors associated with demographic structure for further evaluation.

This study employs an innovative multi-constraint Monte Carlo simulation method to estimate suppressed county-level subgroup cancer counts and extend the downscaling from county to ZIP Code Tabulation Areas (ZCTA). Using cancer incidence data from Utah for the period 2016–2020, the proposed approach is rigorously validated and benchmarked against several machine learning models. The National Cancer Institute (NCI) data and Census data were selected to align with the availability of granular ZIP Code-level cancer data from Utah's SEER cancer registry. Although SEER data are not publicly accessible, their

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: fwang@lsu.edu (F. Wang), Tracy.Onega@hci.utah.edu (T. Onega).

https://doi.org/10.1016/j.healthplace.2024.103411
Received 22 October 2024; Received in revised form 17 December 2024; Accepted 30 December 2024
Available online 6 January 2025



* indicates total volume constraints applied.

Fig. 1. Evolution in population downscaling methods.

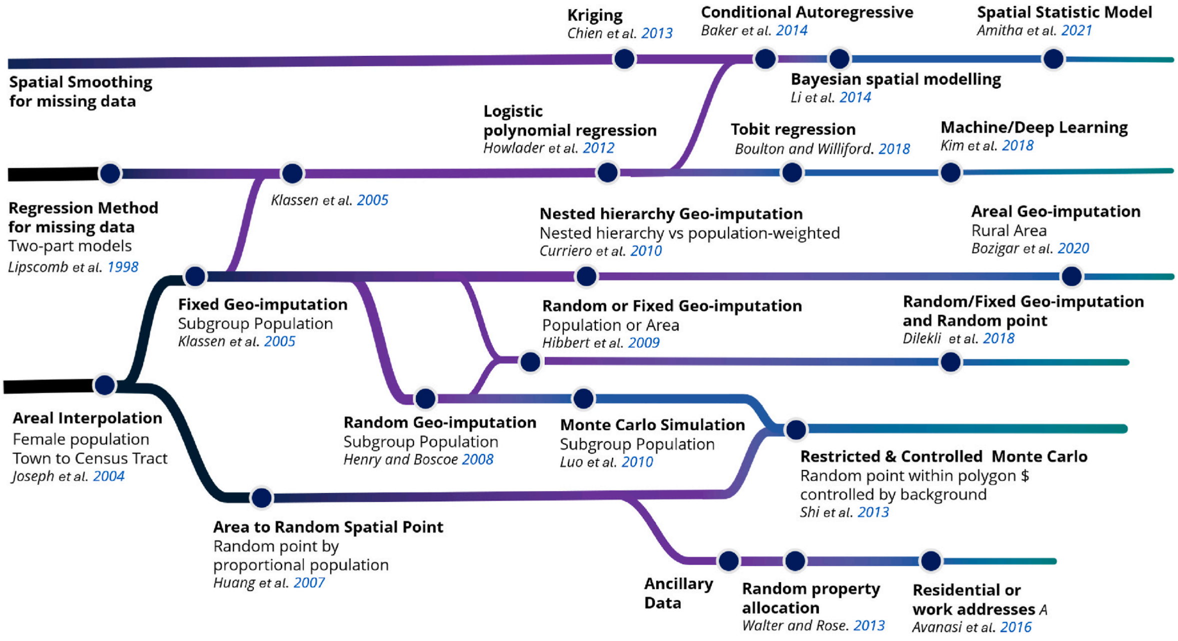


Fig. 2. Evolution of geo-imputation methods for cancer data.

completeness and granularity at the ZIP Code level provide a unique opportunity to evaluate the accuracy and robustness of the Monte Carlo simulation method. This level of validation is particularly critical for estimating suppressed cancer counts in small subgroups, where suppression rules are most frequently applied.

2. Literature review

Methods for addressing missing data and *population downscaling* in health research largely derive from general population science (Chang et al., 2014). Wright introduced dasymetric mapping in 1936 to create downscaled density map based on inhabited and uninhabited areas in towns (Wright, 1936). Subsequent development in this field can be broadly classified into areal interpolation and spatial smoothing, based on whether spatial autocorrelation is considered. Fig. 1 illustrates the knowledge tree structure of related methods.

Initially, *areal interpolation* methods primarily relied on area-weighted calculations within and across administrative boundaries (Goodchild et al., 1980). By incorporating more ancillary data such as cadastral data, land use information, and street data, population-weighted methods have made significant improvements over areal interpolation methods (Xie, 1995; Fisher et al., 1996; Reibel et al., 2007; Comber et al., 2008; Bentley et al., 2013). High-quality ancillary data provide proxies for population downscaling based on geographic boundaries or locations, marking a substantial shift towards data-driven and precise mapping techniques (Liu et al., 2018). Regression methods have been used to integrate the effects of various geographic and demographic variables (Harvey, 2002; Cromley et al., 2012; Harris et al., 2005). The introduction of machine learning and deep learning techniques adds a new dimension to population mapping, enables the incorporation of complex nonlinear relationships, and thereby enhances prediction accuracy (Li et al., 2011; Zhu et al., 2020; Doshi et al., 2023).

Tobler's Pycnophylactic interpolation method was among the first *spatial smoothing* methods for considering spatial autocorrelation and meeting the total number of observations (Tobler, 1979). Subsequent developments have shifted away from total volume constraints towards pure spatial smoothing techniques such as centroid kernel density methods, inverse distance weighting (IDW), geostatistics, and kriging (Martin, 1989; Kyriakidis et al., 2005). Methods like geographically weighted regression (GWR) and spatial Bayesian models integrate spatial correlation with regression and smoothing concepts, and become

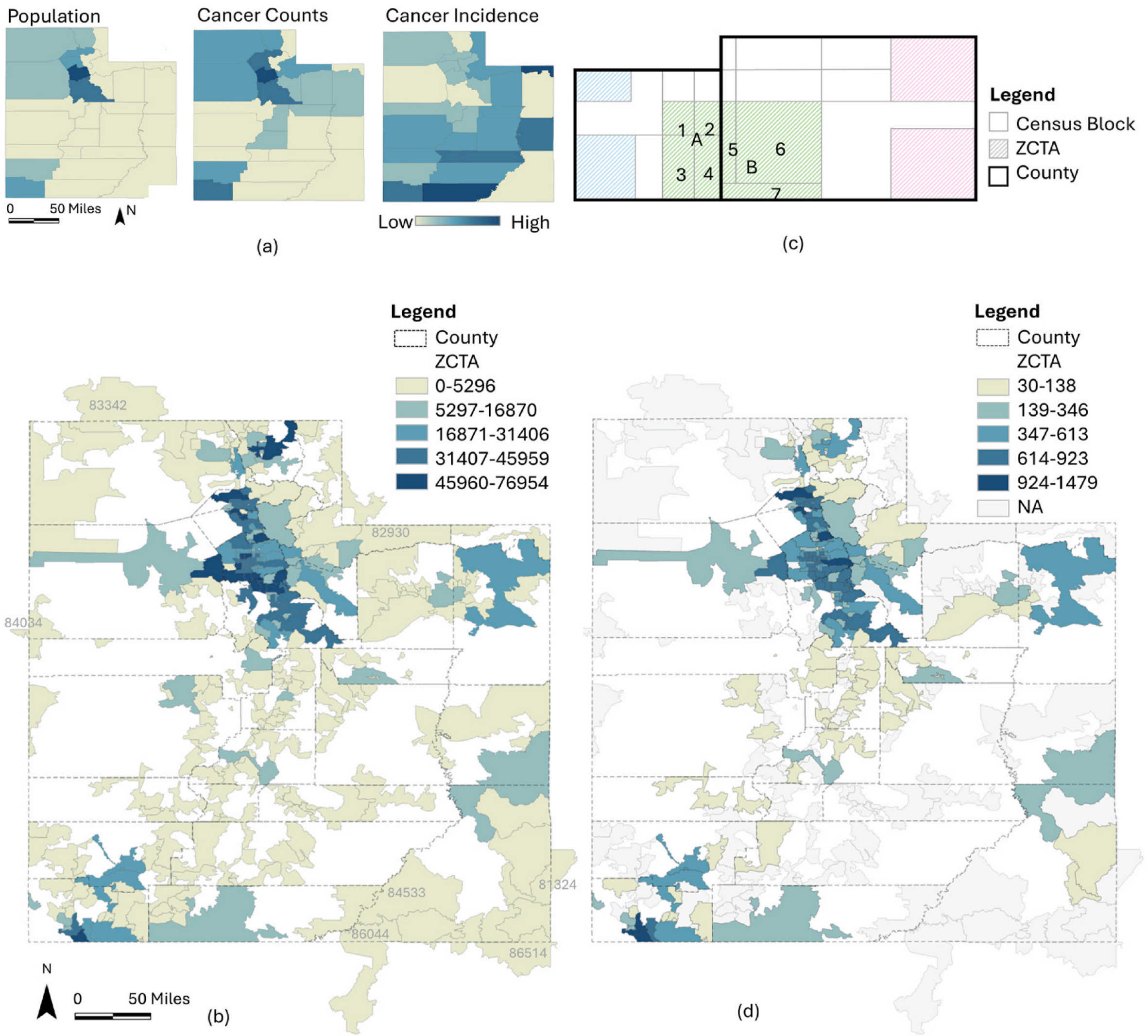


Fig. 3. (a) population and cancer data at the county level data, (b) population at ZCTA level, (c) population interpolation for ZCTA across multiple counties; and (d) Total cancer counts of Utah Registry data in 2016–2020.

an adaptable tool for areal interpolation or spatial smoothing (Liu et al., 2008; Schroeder et al., 2013; Li et al., 2020). By integrating deep learning and machine learning methods with geographic features, Geospatial AI (GeoAI) represents the forefront in a knowledge tree for spatially varying models on population downscaling (Liu, 2024).

These methods need to be adjusted to account for the specific nature of medical data and enforce total volume constraints (Lam, 1983), when applied to cancer data processing. More recently, areal interpolation has gained more attention than spatial smoothing in health research and been expanded into *geo-imputation* (Chien et al., 2013). Fig. 2 illustrates the evolution of the related methods suitable for cancer data imputation.

Areal interpolation method was first introduced to interpolate breast cancer cases from towns to census tracts based on the proportion of females (Joseph Sheehan et al., 2004). The methods are termed "geo--imputation" by using population data segmented by sex, age, and race to allocate cancer proportions to smaller spatial unit centroids (Klassen et al., 2005). Based on comparisons of multiple imputation methods in nested geographic units, population weighting has been validated as more optimal than area weighting (Curriero et al., 2010). However, a recent study argued that population weighting could underestimate case numbers in rural areas (Dilekli et al., 2018). These discussions highlight the importance of selecting appropriate auxiliary data for downscaling to better align with the target dataset.

Random cancer case simulation was introduced in geo-imputation to conduct uncertainty analysis on distribution statistics, and thus extended fixed proportion allocation to probability distribution (Henry et al., 2008). Building on this, the Monte Carlo simulation method was further applied with simple constraints to improve dynamic proportion constraints during the case generation process (Luo et al., 2010). Random spatial imputation began with generating random spatial points based on population proportions to enhance data granularity (Huang et al., 2007). These methods later incorporated dasymetric mapping techniques by using additional ancillary data such as property locations and daily activity spaces for finer spatial distribution (Walter et al., 2013; Avanasi et al., 2016). A combination of random case numbers and random geographic locations was proposed as the *Restricted and Controlled Monte Carlo method,* which combines the restriction of geographic boundaries and controls of health data prevalence by population (Shi et al., 2013).

Similar to population downscaling, regression methods and spatial smoothing were also applied on cancer data. For example, some employed both geo-imputation and regression methods to disaggregate or predict cancer cases at different scales (Klassen et al., 2005). Subsequent regression and prediction methods, ranging from Tobit models to machine learning, were widely applied to various types of data imputation (Lipscomb et al., 1998; Boulton et al., 2018; Kim et al., 2018). Spatial smoothing methods such as kriging, spatial Bayesian, and spatial panel models were also commonly used (Chien et al., 2013; Baker et al.,

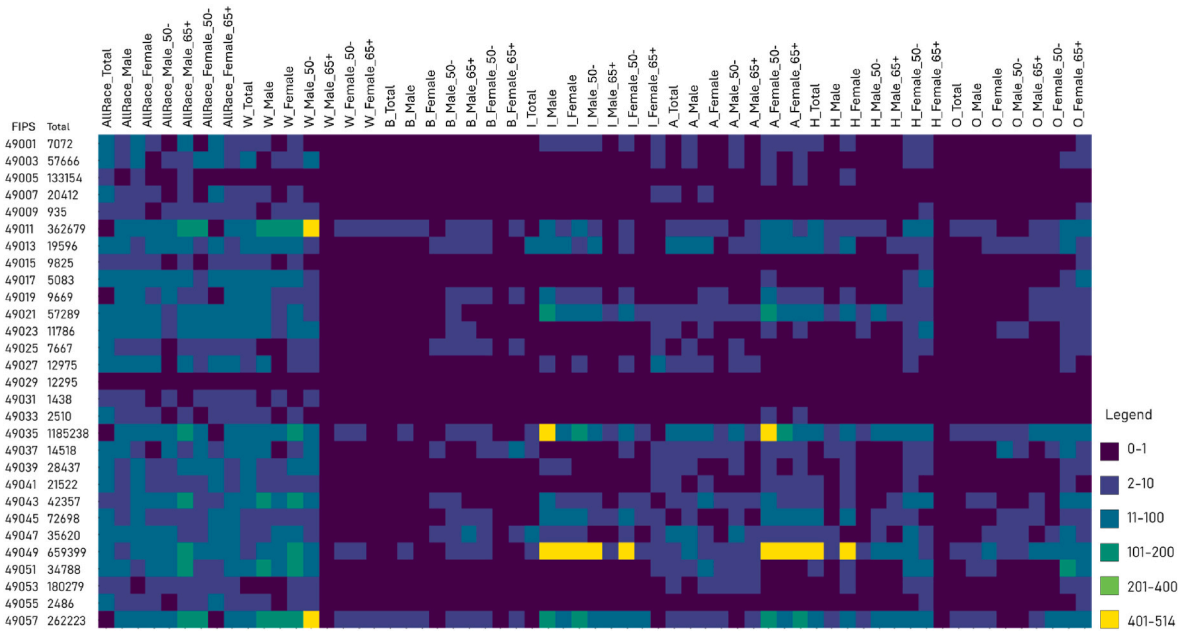


Fig. 4. Population discrepancy between county totals and ZCTA-interpolated values.

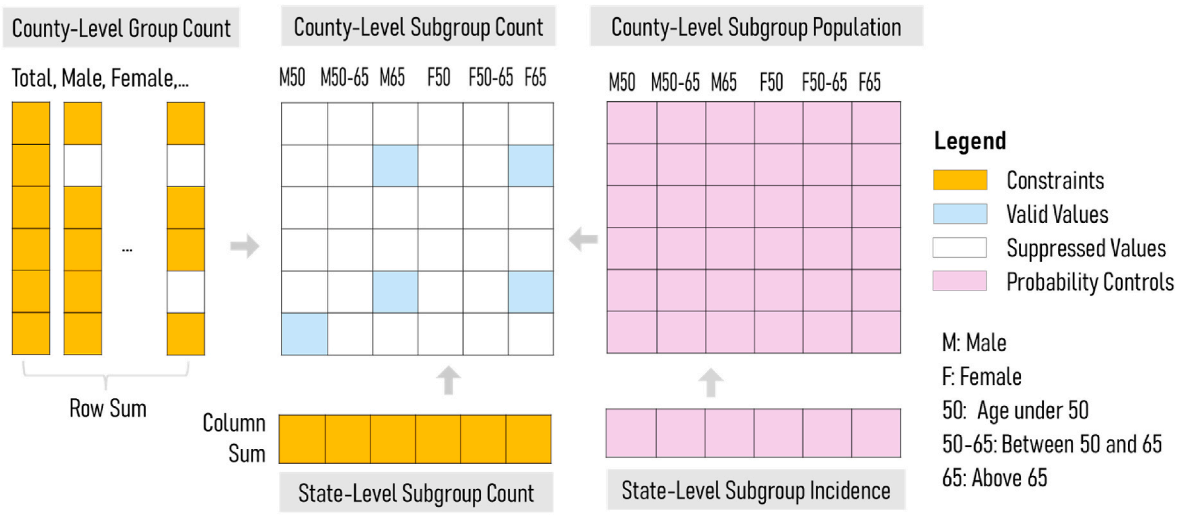


Fig. 5. Multiple constraints in Monte Carlo simulation.

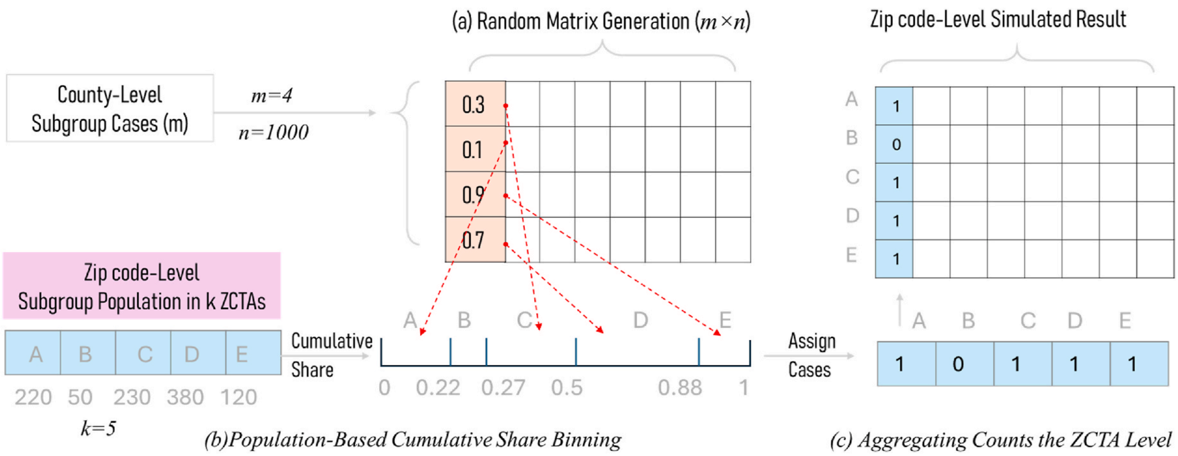

(a) Random Matrix Generation ($m \times n$)
County-Level Subgroup Cases (m)
$m=4$
$n=1000$
0.3
0.1
0.9
0.7
Zip code-Level Subgroup Population in k ZCTAs
A B C D E
220 50 230 380 120
$k=5$
Cumulative Share
A B C D E
0 0.22 0.27 0.5 0.88 1
(b)Population-Based Cumulative Share Binning
Zip code-Level Simulated Result
A B C D E
1
0
1
1
1
Assign Cases
A B C D E
1 0 1 1 1
(c) Aggregating Counts the ZCTA Level

Fig. 6. Monte Carlo Simulation for estimating ZCTA subgroup cases.

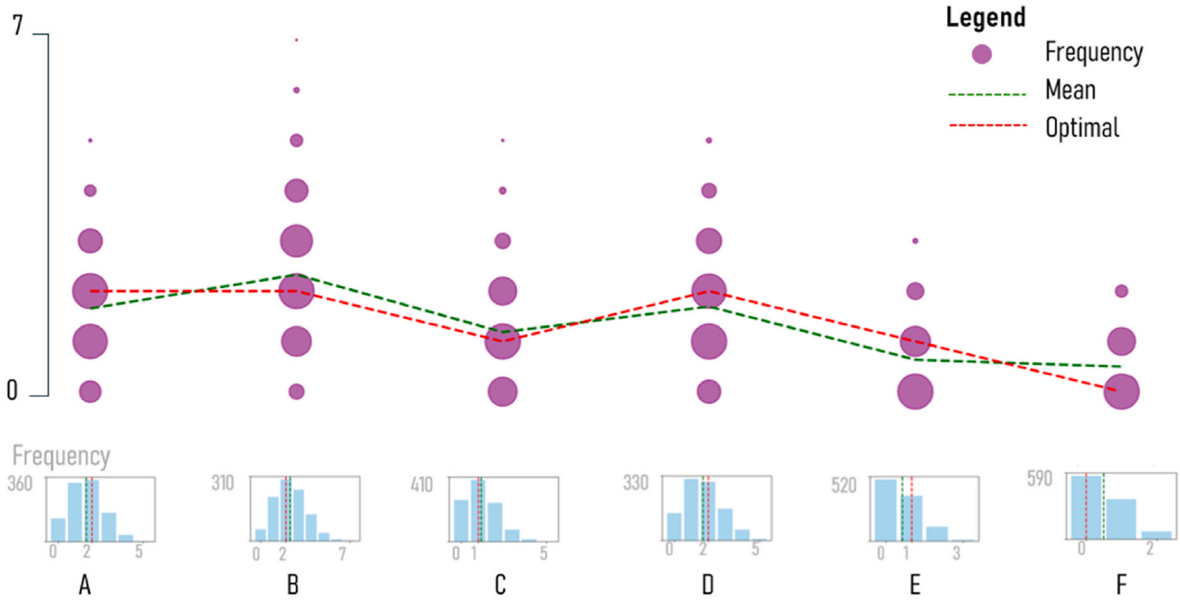


Fig. 7. An illustrative example for simulating suppressed cancer counts in 6 ZCTAs: the top shows optimal scenario vs. means, and the bottom shows frequency histograms.

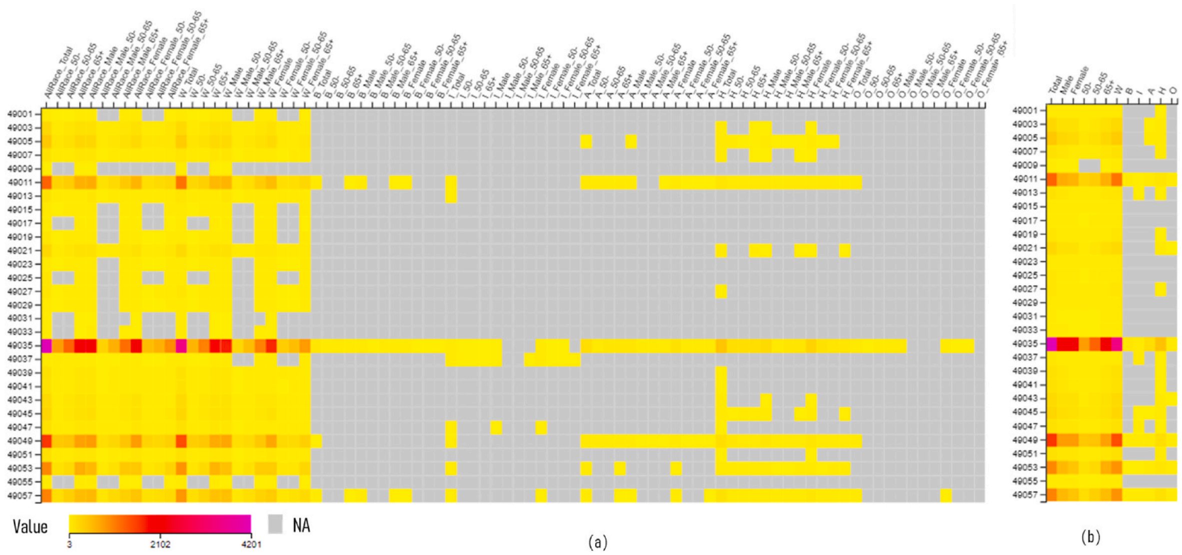


Fig. 8. Cancer Incidences at the county level in Utah: (a) NCI data, and (b) UCR data.

Table 1
Annual cancer incidences in ZCTAs based on the UCR data.

Population Groups	# of valid observations	Min	Max	Mean	Standard deviation
Total	153	6.0	295.8	72.6	62.5
Male	153	2.6	162.8	37.9	32.7
Female	153	2.2*	133.0	34.7	30.2
Under 50	119	2.2*	52.6	17.2	11.9
50–64	119	2.8	81.0	26.1	17.0
65+	153	2.4	213.6	38.2	35.2
White, Non Hispanic (W)	153	2.4	274.0	63.1	54.5
Black, Non Hispanic (B)	10	2.2*	4.8	3.3	1.0
American Indian or Alaskan Native, Non Hispanic (I)	4	2.4	5.2	3.8	1.3
Asian or Pacific Islander, Non Hispanic (A)	47	2.2*	16.2	4.5	2.9
Hispanic, Any Race (H)	85	2.2*	36.8	10.4	7.9
Others (O)	7	2.2*	7.0	3.8	2.0

Note: * 2.2 = 11 (i.e., threshold for data suppression)/5 (i.e., number of years for 2016–2020).

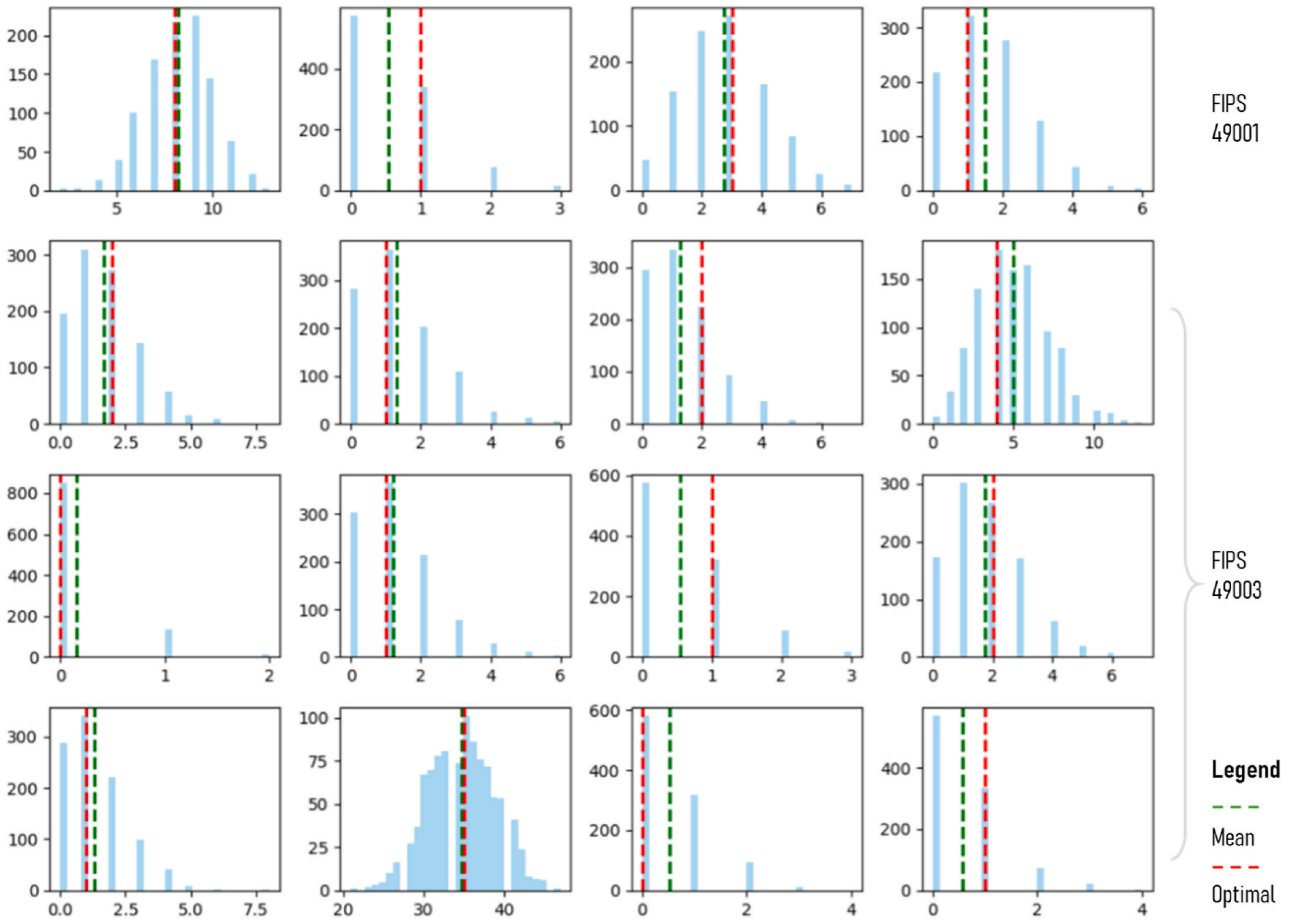


Fig. 9. Simulations for Cancer Incidences of White Male 65+ Subgroup across 16 ZCTAs in Two Counties in Utah (x-axis for Cancer Count, y-axis for Frequency among 1000 Scenarios).

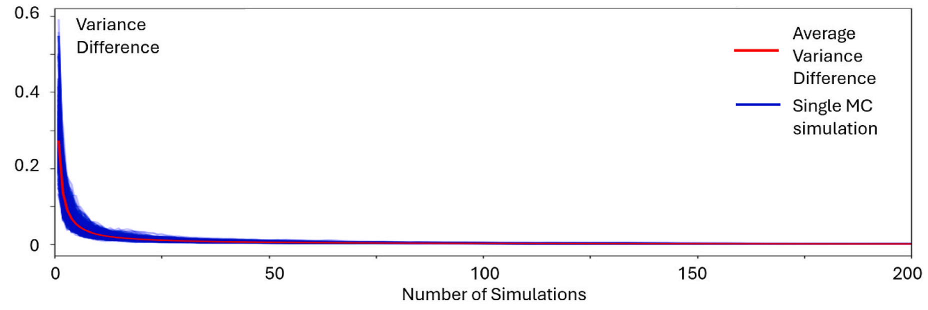


Fig. 10. The variance difference between the cumulative mean and the overall mean with the number of iterations (n).

2014). However, these methods cannot ensure total volume consistency.

In general, despite significant advancements in data imputation and population downscaling techniques, several critical gaps remain unaddressed in the context of health data, particularly cancer data. Most existing methods, such as regression-based approaches and machine learning models, struggle to simultaneously satisfy demographic and geographic constraints, often leading to biased estimates. For example, population-weighted methods, while effective in some scenarios, have been shown to underestimate rural case numbers and overestimate urban cases, reducing their robustness across diverse geographic contexts. Furthermore, many spatial smoothing and regression methods fail to ensure total volume consistency, a critical requirement for accurate health data imputation.

The Restricted and Controlled Monte Carlo method generally provides accurate and comprehensive insights for cancer data imputation while preserving total volume at higher geographic scales (Hu et al., 2015; Huang et al., 2022). However, additional challenges remain in implementing simulations that span nested geographic levels and account for multiple subgroup population controls. This study addresses these challenges by proposing an innovative multi-constraint Monte

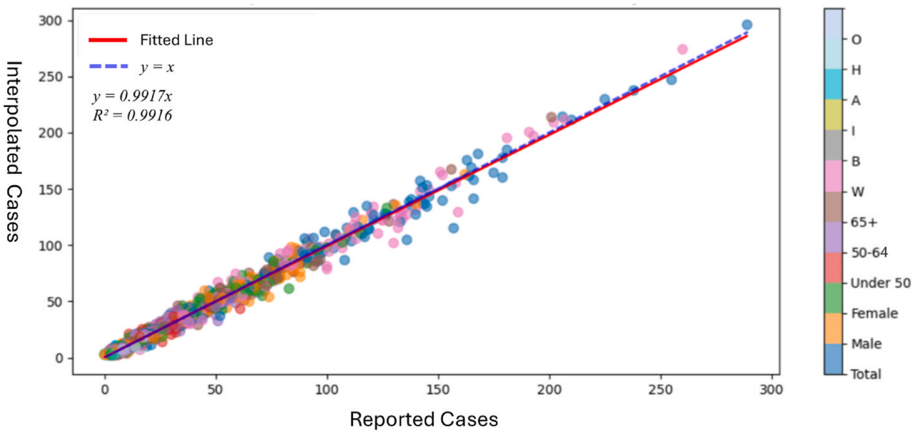


Fig. 11. Reported vs. interpolated cancer incidences by the Monte Carlo Method.

Table 2
Confusion matrix for interpolating cancer incidences across ZCTAs.

		UCR Values	
		Suppressed (<2.2)	Reported (≥2.2)
Interpolated values	$x < 2.2$	True Positive: 2110	False Negative: 48
	$x \geq 2.2$	False Positive: 202	True Negative: 1108
Total number of values		2312	1156
Precision rate		91.3%	95.8%

Table 3
Performance comparison across interpolation methods.

Methods	Mean Absolute Error (MAE)*	Correlation Coefficient (r)*	Precision % based on Confusion Matrix**
OLS	27.95	0.827	72.8%
GLM	121.10	0.606	46.9%
RF	12.57	0.480	41.0%
GBM	12.49	0.606	33.3%
XGB	10.29	0.910	33.3%
Monte Carlo	*3.49*	*0.991*	*92.3%*

Notes: * Number of observations (UCR reported values) $n = 1156$; ** Percentage of interpolated values in the right categories of suppression.

Carlo simulation method. The method ensures total volume consistency, robust performance across urban and rural contexts, and reliable estimation of suppressed cancer counts at fine geographic levels, thus filling a critical gap in the existing literature. By integrating dynamic demographic constraints and probabilistic case generation, the proposed approach provides actionable insights for precision public health interventions.

3. Data and methodology

3.1. Data sources and preprocessing

3.1.1. Open datasets for MC cancer data downscaling

Annual average statistics of cancer incidence counts by age, sex, and racial/ethnic group: The data for Utah, spanning the years 2016–2020,

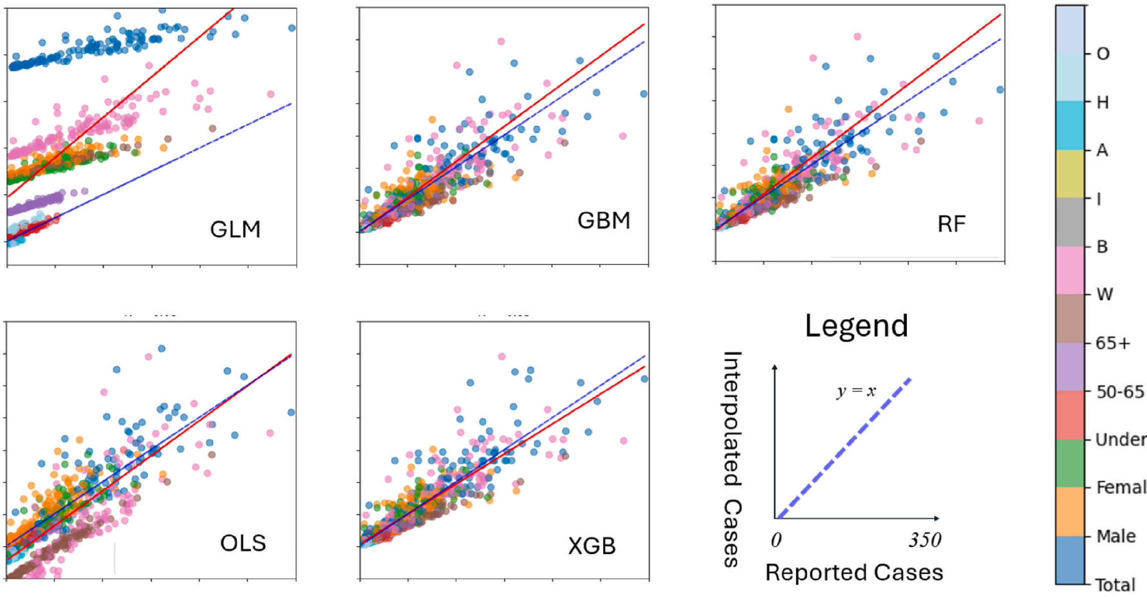


Fig. 12. Reported vs. interpolated cancer incidences by machine learning models.

was obtained from the National Cancer Institute (NCI) database (https://statecancerprofiles.cancer.gov/incidencerates/). It is categorized by race and ethnicity (All Races, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic American Indian and Alaska Native, Non-Hispanic Asian and Pacific Islander, Hispanic, Other), sex (female, male), and age groups (under 50, 50–65, over 65). There are a total of 42 subgroups, i.e., (1 all races + 6 racial-ethnic categories) × 2 sex categories × 3 age groups. Removing "all races" results in 36 meaningful subgroups, i.e., 6 racial-ethnic categories × 2 sex categories × 3 age groups. Both the state and county level data are utilized, where the state-level aggregate values will be used as constraints for Monte Carlo (MC) simulations to impute missing values at the county level. These datasets comply with established privacy protection standards, particularly through the suppression of annual avareage counts fewer than 3 cases over five years, effectively preventing the direct exposure of individual-level data. Fig. 3a shows the total population, cancer counts and incidence at the county level in Utah.

Population data by age, sex, and racial/ethnic group: It was extracted from the 2020 Decennial Census data on Demographic and Housing Characteristics (https://api.census.gov/data/2020/dec/dhc.html). These data are categorized by race and ethnicity using the P12 series, which includes the following groups: All Races (P12), Non-Hispanic White (P12I), Non-Hispanic Black (P12J), Non-Hispanic American Indian and Alaska Native (P12K), Non-Hispanic Asian and Pacific Islander (P12L), and Hispanic (P12H). Each racial and ethnic category is further disaggregated by sex (female, male) and detailed age groups. For the purpose of this study, the age groups are aggregated into three broader categories: under 50, 50–65, over 65. It was then aggregated to the corresponding 42 population subgroups consistent with the above cancer data at the state, county, and ZIP Code Tabulation Areas (ZCTA) levels in order to match with cancer data. Such data were used for Monte Carlo simulations to calculate missing case counts from the state to the county level, and from the county to the ZCTA level.

Note that a ZCTA may span across different counties or even states, as shown in Fig. 3b. For better accuracy, population in each portion of such a ZCTA is calculated using census block data as auxiliary data. This approach utilizes the clear spatial correspondence between ZCTAs and the census blocks they encompass, eliminating the need for interpolation. Cross-county ZCTAs are subdivided into county-level subregions based on their spatial relationships. Population counts for each subregion are then calculated by aggregating data from the corresponding census blocks. These population estimates serve as the basis for allocating cancer counts to each county. After the county-level allocations are finalized, the subregions are recombined to produce a comprehensive dataset of cancer counts for the entire ZCTA. For example, as shown in Fig. 3c, the green ZCTA is split between two counties. Population in portion A is the sum of four blocks on the left side, and population in portion B is the sum of three blocks on the right side. Block is the smallest census unit with population data by subgroups. Therefore, cancer data at the county level will be disaggregated to whole or partial ZCTAs within each county according to the underlying demographic structure, and interpolated cancer data in partial ZCTAs will then be consolidated to whole ZCTAs.

As shown in Fig. 3c, some areas with very small population are not covered by ZCTAs, and thus are left out. As a result, the aggregated population from the ZCTA level in a county may be slightly below its total population. As shown in Fig. 4, the discrepancy ranges between 0 and 500 across 29 counties in Utah, with an average error rate of 0.011%.

3.1.2. Utah Cancer Registry data for model validation

Population-based cancer incidence counts were taken from the Utah Cancer Registry (UCR) (https://uofuhealth.utah.edu/utah-cancer-registry), which has been providing high quality cancer data since 1968. The UCR is a Surveillance, Epidemiology, and End Results (SEER) registry (https://seer.cancer.gov/), which requires adherence to the highest quality standards for ascertainment and reporting. The *Utah Cancer Registry (UCR) dataset* comprises 57,534 cancer cases diagnosed over the five-year period from 2016 to 2020, geocoded to county and ZIP code levels. This dataset, especially at the ZIP code level, provides an opportunity for validating our cancer data imputation (Fig. 3d). The same minimum number of cases (11) is applied in cancer data suppression at both county and ZIP code levels. In other words, the numbers for cancer cases of fewer than 11 are suppressed for any demographic groups or total at either the county or ZIP code level. The demographic subgroups correspond to those NCI data as discussed earlier. Use of data from the UCR was approved by the Institutional Review Board (IRB) at the University of Utah. This study does not directly handle suppressed values; rather, they are utilized to validate the accuracy of the model's results. Specifically, the results generated through Monte Carlo and geo-imputation methods are evaluated using numeric scores for valid values and binary scores for suppressed values, ensuring robustness and reliability.

One issue encountered in data processing involves some minor discrepancy between ZCTA area codes in the NCI data and ZIP codes in the UCR. Out of 323 entries in the UCR data, 289 ZIP codes match with ZCTA codes. For the remaining 34 ZIP codes, only one (ZIP Code 84068) contains valid data. To resolve this issue, we utilize USPS services to query the city name, geocode the address of these ZIP codes, and assign them to their corresponding ZCTAs (https://tools.usps.com/zip-code-lookup.htm?citybyzipcode). For instance, postal ZIP code 84068 is assigned to ZCTA 84060, corresponding to Park City, Utah.

3.2. Methods

As stated previously, compared to regression models or machine learning, the Monte Carlo method excels in achieving total consistency through explicit multi-constraints, such as population subgroups and cancer incidence rates. While traditional methods offer flexibility, they often fail to satisfy both total constraints and spatial consistency. The Monte Carlo approach also reduces bias when estimating suppressed values for small samples by leveraging probabilistic modeling, making it well-suited for spatially constrained health data. Our method involves two steps: (1) a Multi-Constraints Monte Carlo (MC) simulation to estimate suppressed cancer counts at the county level, (2) a Monte Carlo with Population Cumulative Share approach to downscale from counties to ZIP Code Tabulation Areas (ZCTAs). An Optimal Scenario Selection strategy is used to refine simulation parameters for improved accuracy. To benchmark our approach, we compare it against machine learning models, testing cases with population data alone, with additional ancillary variables (e.g., poverty, education), and using cancer incidence as the target variable.

3.2.1. Multi-constraints MC for suppressed county cancer counts

The multi-constraint MC simulation method is proposed to estimate suppressed values of county-level subgroup cancer counts. Fig. 5 illustrates the constraints and the task to be accomplished. The constraints are reported cancer counts of county-level major demographic groups (e.g., Total, by 2 sexes, 3 age groups, 6 racial-ethnic groups) and state-level subgroups (e.g., by the intersected subgroups of sex, age and race-ethnicity). This method uses a combination of the county-level subgroup population distribution and state-level cancer incidence rates to control the probability in the Monte Carlo simulation process to estimate the suppressed county-level cancer counts in the intersected subgroups. It not only ensures consistency across various data levels but also accounts for demographic structures that drive varying cancer risks.

The Monte Carlo method is divided into four steps.

(1) Extract constraints to construct an initial data framework for row sums and column sums from state-level subgroup counts and county-level group counts. Only the rows containing suppressed values will be simulated.

(2) Randomly assign 1 case to the cells according to the proportion of each subgroup in the total population weighted by state-level incidence, ensuring accuracy in the representation of subgroup proportions throughout the data filling process. Once assigned, subtract 1 from row sum and column sum.
(3) Repeat Step 2 until all row and column constraints are satisfied.
(4) Repeat Steps 2–3 1000 times and choose the optimal simulation scenario based on the mean and standard deviation (detailed in sub-section 3.3).

This method is deployed from groups to subgroups hierarchically, as the prevalence of suppressed data increases with more detailed demographic groupings. For example, it first estimates suppressed cancer counts for "females" in counties, then uses the information to estimate suppressed cancer counts for "females under 50 years", and finally uses newly interpolated counts to estimate suppressed cancer counts for "Black females under 50 years." As a result, the missing cancer data for 36 subgroups (i.e., 2 sexes × 3 age groups × 6 racial-ethnic categories) at the county level are estimated. Logically, the reliability of estimation declines with finer groupings.

To assess the convergence and stability of the Monte Carlo simulation method, a sensitivity analysis is conducted to evaluate the relationship between the number of simulations and the variance difference between the cumulative mean and the overall mean.

3.2.2. MC with Population Cumulative Share for suppressed ZCTA cancer counts

For estimating subgroup cancer counts at the ZCTA level, another Monte Carlo simulation method is designed by leveraging subgroup cancer counts at the county level and subgroup population data for all ZCTA units. This concept is similar to estimating suppressed cancer counts in some county-level subgroups, but we adopt Python's robust capabilities for data manipulation and statistical analysis to ensure high computational efficiency and accuracy.

The method is implemented in four steps.

(1) *Random matrix generation.* For each subgroup, a matrix of random numbers ranging from 0 to 1 is generated to represent potential subgroup distributions across different simulations. This matrix size is $m \times n$, where m represents the number of cases (subgroup counts at the county level) and n is the number of desired simulations (e.g., $n = 1000$). Random numbers are drawn from a uniform distribution $U(0,1)$ to ensure equal likelihood for all initial allocations before applying demographic constraints.
(2) *Population-based cumulative share binning.* The population data for each subgroup is processed to calculate cumulative shares, which define the range of values (bins) for assigning random numbers to specific ZCTAs. For each ZCTA i, the cumulative share is calculated as:

$$C_i = \sum_{j=1}^{i} \frac{P_j}{P_{Total}}$$

where P_j is the population of the subgroup in ZCTA j, and P_{Total} is the total population of the subgroup in all ZCTAs. The resulting cumulative share for each ZCTA defines a range: ZCTA i corresponds to the interval $[C_{i-1}, C_i)$, where $C_0 = 0$. These intervals are subsequently used to create bins for categorizing the random numbers. These bins effectively map the random numbers to specific population segments so that the generated distributions are proportional to actual demographic profiles. Any subgroup with zero population is excluded.

(3) *Aggregating counts in bins at the ZCTA level.* Utilizing the Pandas package function *pandas.cut* to rapidly categorize and aggregate the simulated data into defined bins based on the cumulative share ranges calculated in step 2.
(4) *Result selection.* Similar to the county-level approach, the optimal simulation scenario is chosen, based on criteria that minimize the variance between the simulated and observed county-level counts. The simulation iteration that most closely matches the observed distribution is selected as the result. Details on the selection process are provided in the next sub-section.

Fig. 6 uses an illustrative example to explain the process. In a county with m (=4) cases in a subgroup, a random matrix is generated with each column representing one scenario out of n (=1000) simulations. A series of 4 numbers (0–1) are generated in scenario 1 (1st column highlighted in Fig. 6a). Based on the subgroup population across k (=5) ZCTAs, a cumulative share bin is constructed with each segment length representing its proportion (probability) out of the total subgroup, and a match is made by assigning each random number in column 1 to their corresponding segment (Fig. 6b). As a result, the four random numbers (0.3, 0.1, 0.9 and 0.7) are assigned to segments corresponding to four ZCTAs (C, A, E and D).[1] Therefore, these four ZCTAs receive 1 case each and B receives none, and the allocation of 4 cases is completed. The process repeats 1000 times.

3.2.3. Selecting optimal scenario

A large number (here 1000) of independent Monte Carlo simulations are executed for each disaggregation. Various criteria may be used to assist the selection of an optimal simulation scenario (Zhao et al., 2024). For each area, the mean value across all simulation scenarios approximates the most likely choice. However, such a choice does not satisfy the constraints of cancer counts at a higher geographic level or for given demographic groups. Therefore, the minimum variance as a common and robust measure of data dispersion is chosen as the selection criterion (Buzaianu et al., 2017; Gupta et al., 1962). The simulation with the minimum standard deviation from the means indicates the closest alignment to the average pattern and is thus the best representation of the possible subgroup distribution.

Here the formulation uses the disaggregation of a subgroup from counties to ZCTAs as an example. Let X_i^k represent the simulated value list obtained from the k-th simulation for an estimated ZCTA (i), calculate the average estimated count across the n simulations, such as $\overline{X}_i = \frac{1}{n}\sum_{k=1}^{n} X_i^k$, where X_i^k is the value at row i and column k of simulation matrix.

For each simulation k, compute the deviation of its standard deviation σ_k from the mean simulation $\overline{X}_i$, such as $\sigma_k = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(X_i^k - \overline{X}_i\right)^2}$, where m is the total number of ZCTAs, or the row number in the matrix X. The simulation with the smallest σ_k is selected as the result.

Fig. 7 illustrates a county of six ZCTAs. Its top highlights how close the chosen (optimal) scenario is from the means across the six ZCTAs, and the bottom shows the frequency histograms for estimated suppressed values at 1000 times.

3.2.4. Machine learning methods for comparison

To evaluate the performance of the proposed Monte Carlo simulation method, we conducted a comparative analysis using four popular machine learning models—Generalized Linear Model (GLM), Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGB)—along with the traditional Ordinary Least Squares (OLS) regression model. These models are utilized to estimate suppressed cancer counts at the ZIP code level for subgroup populations. GLM extends traditional linear regression by allowing the response variable to follow non-normal distributions through link functions, offering flexibility for various types of outcomes (e.g., binary, count). Its

[1] It is possible for multiple random numbers to fall in one segment along the cumulative share bin. In that case, the corresponding ZCTA is assigned multiple cancer cases.

interpretability makes it a robust baseline, though it may struggle with non-linear relationships. OLS regression serves as a classic linear modeling approach, assuming a linear relationship between predictors and the response variable. While straightforward and interpretable, it shares GLM's limitation in handling non-linear patterns. Random Forest (RF), GBM, and XGB are ensemble learning methods that are well-suited for modeling complex, non-linear relationships. RF aggregates predictions from multiple decision trees, while GBM and XGB use boosting techniques to sequentially refine model accuracy. XGB, as an advanced implementation of GBM, is optimized for computational efficiency and regularization to reduce overfitting.

To supplement the primary population and cancer data, we incorporate additional demographic and socioeconomic variables from the American Community Survey (ACS). These variables include total population estimates for educational attainment, income levels, and health insurance coverage, as well as detailed subsets such as the population with a bachelor's degree (B15003), individuals with income below the poverty level in the past 12 months (B17017), and those 65 years and over with one type of health insurance coverage (B27010). The county-level ACS data are utilized for machine learning model training, leveraging a broader dataset to enhance model robustness and capture regional variations. In contrast, ZCTA-level data for Utah are reserved exclusively for model testing, ensuring an independent evaluation of the model's ability to estimate suppressed cancer counts in smaller geographic units.

This study focuses on accurately predicting actual case counts, using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as the key evaluation metrics for model comparison. MAE measures the average deviation between predicted and observed values, offering a straightforward assessment of absolute prediction accuracy. RMSE, which emphasizes larger errors due to squaring deviations, is particularly important for addressing significant errors, such as those with major influences on cancer case predictions. Mean Absolute Percentage Error (MAPE), which expresses errors as a percentage of actual values, is useful for small populations. However, MAPE is not applied in this study due to the use of primarily non-suppressed validation data.

4. Results

4.1. Simulation results

The NCI county-level cancer data for the 29 counties in Utah are used in the simulations.[2] The suppressed county-level subgroup cancer counts are estimated by choosing the optimal scenarios among the large number of simulation rounds. Such data in combination with reported cancer counts at the county level are subsequently used in interpolating ZCTA cancer counts. The ZCTA cancer data are not available from the NCI but are provided from the Utah Cancer Registry (UCR) after the data suppression rule is enforced. Therefore, the UCR data at the ZCTA level are used for validating the interpolated ZCTA data.

Our comparison of the county-level data from the two sources reveals a high consistency between the two. However, there are some differences between the two in terms of availability of values. While covering the same five years 2016–2020, the NCI county-level data are the *annual average* cancer incidences over the five years by suppressing values < 3 (i.e., <15 for the five-year totals), whereas the UCR county-level data are the *five-year total* cancer incidences by suppressing values < 11. Therefore, a small number of suppressed counts in the NCI data are available in the UCR. The NCI county-level data include cancer incidences by 84 demographic groups/subgroups in 29 counties, thus $29 \times 84 = 2436$ units, of which 783 are valid. As shown in Fig. 8a, by columns, data on 'All races' and 'White' exhibit fewer suppressed values than other subgroups; and by rows, counties with larger numbers of population have fewer missing values. As shown in Fig. 8b, the 2016–2020 UCR data only contain 12 major demographic groups, therefore $29 \times 12 = 348$ units, with 258 valid.

For consistency with the NCI data, we convert the 2016–2020 UCR data to annual average cancer incidences (and thus with decimal points). Table 1 provides basic statistics for the ZCTAs based on the UCR data. Note that only reported (unsuppressed) units are included, and thus the minimum values for several population groups are 2.2 (i.e., 11/5 or the threshold for data suppression divided by the number of years). The cancer incidences for males, individuals over 65, and Whites are notably higher. In addition to all population ("Total"), cancer data for population subgroups includes six racial-ethnic subgroups (W, B, I, A, H, O), two sexes (Male and Female), and three age groups (50, 50–65, 65+). As stated previously, the 2020 Census Demographic Profile covers 299 ZIP code areas, and the UCR data includes 323 ZCTAs. After merging the two and eliminating ZCTAs with no valid data, 289 ZCTA are retained. For 12 major population groups, there are a total of $289 \times 12 = 3468$ units, 1156 of which have valid values that can be used for data verification.

As explained in subsection 3.1, the study begins with implementing constrained simulations for imputation of missing values at the county level. After completing the data fill for the "Total" and "White" categories, we simulated other racial-ethnic groups such as "Black," "Indian," "Asian," "Hispanic," and "Others." We then simulated data for the three age groups and the two sex groups to fill in more data gaps. The process continued until the simulations for the 42 intersected subgroups of race-ethnicity (7), age (3) and sex (2) were completed.

For illustration, the simulation results for one subgroup (white male age 65+) in two counties for a combination of 16 ZCTAs are summarized in Fig. 9. Generally, most values follow a normal distribution pattern. When values are comparatively low, there is a noticeable biased distribution. Overall, the optimal values, chosen based on the minimum standard deviation, demonstrate good consistency with the mean values.

4.2. Sensitivity analysis of Monte Carlo Simulations

To evaluate the stability and convergence of the Monte Carlo simulation results, we conduct a sensitivity analysis based on the variance difference between the cumulative mean and the overall mean of all simulations. This analysis assesses how the simulation output stabilizes as the number of iterations increases. The sensitivity analysis involves $N = 1000$ Monte Carlo simulation iterations by repeating 1000 times with random shuffling of the simulated data in each repetition. For each iteration n, the cumulative mean of the simulated values up to iteration n is calculated and compared to the overall mean across all iterations. The variance of these differences is computed at each step, resulting in a metric that reflects the stability of the simulations.

Fig. 10 shows the variance difference between the cumulative mean and the overall mean as a function of the number of iterations (n). Each blue line represents a single repetition, while the red line depicts the average variance difference across all repetitions. The variance difference decreases rapidly within the first 50 iterations and plateaus as n approached 200. This rapid convergence demonstrates that the Monte Carlo method achieves stability early in the simulation process, with diminishing returns for additional iterations beyond this point. Despite minor variability among the 1000 repetitions, as illustrated by the blue lines, the overall trend remains consistent. The red line, representing the average variance difference, confirms that the variance difference stabilizes quickly and uniformly across repetitions. The results highlight the robustness of the Monte Carlo approach, showing that a relatively small number of iterations (approximately 50) is sufficient to achieve stable and reliable results. Beyond this threshold, additional iterations have a minimal impact on reducing variance differences, underscoring

[2] Visit the website: https://www.statecancerprofiles.cancer.gov/incidencerates/index.php?stateFIPS=49&areatype=county&cancer=001&race=00&sex=0&age=001&type=incd&sortVariableName=rate&sortOrder=default&output=0#results.

the computational efficiency of the method.

The sensitivity analysis provides critical insights into the convergence behavior of the Monte Carlo simulations, ensuring that the estimates are stable and reliable. This rapid convergence is particularly advantageous for large-scale applications, where computational efficiency is essential. These findings confirm the suitability of the Monte Carlo framework for estimating suppressed cancer counts under the constraints of spatial and demographic data.

4.3. Validation and comparison with machine learning models

As stated in section 4, the validation is assessed at the ZCTA level. Specifically, the interpolated cancer incidences by the Monte Carlo simulation method are compared to the corresponding incidences as reported in the UCR data. For the 12 major population groups across 289 ZCTAs, among a total of $289 \times 12 = 3468$ possible units, 1156 reported values can be used for data verification and 2312 are suppressed.

Fig. 11 shows the distribution of two series of data (interpolated vs. reported) across 1156 observations. The regression model has a slope (0.992) close to 1 with $R^2 = 0.992$ and correlation coefficient $r = 0.996$, reflecting very high predictive accuracy. Another important measure of performance is Mean Absolute Error (MAE) = 3.407.

Another validation is assessed on the complete set of 3468 entries across ZCTAs including 1156 reported values and 2312 suppressed values. The assessment examines whether the interpolated values fall in the right category of suppression decision. If an interpolated value x for a suppressed unit is below the suppression threshold set at 2.2, i.e., $x < 2.2$, it is considered a valid interpolation and thus coded “True Positive”. If an interpolated value x for a suppressed unit is equal to or larger than 2.2, i.e., $x \geq 2.2$, it is considered an invalid interpolation and thus coded “False Positive”. Similarly, for reported values, if an interpolated value $x < 2.2$, it is considered an invalid interpolation and coded “False Negative”; if $x \geq 2.2$, it is a valid interpolation and coded “True Negative”. The result is summarized in Table 2, among the 2312 suppressed entries, 215 are interpolated above the suppression threshold, resulting in a precision rate of 91.3%; among the 1156 reported values, only 52 are below the suppression threshold, and thus a precision rate of 95.8%. So the overall precision rate is 92.8%, i.e., (2110 + 1108)/(2312 + 1156).

Note that some values above the suppression threshold are suppressed in order to prevent the mathematical derivation of other below-threshold values. For example, if the cancer incidences in a ZCTA across three age groups (<50, 50–64, 65+) are 3, 12 and 28 for a total of 43, the count for the age group of 50–64 needs to be suppressed in order to prevent one from deriving the count for the age group <50 as 43-28-12 = 3. Such an “over-suppression” practice inflates the number of False Positive and artificially brings down the precision rate. That helps explain the relatively low precision rate of 91.3% on the suppressed samples. That is to say, the overall precision rate of the Monte Carlo method is higher than 92.8% as reported here.

One concern raised in the literature review is that population weighting interpolation methods could underestimate case numbers in rural areas (Curriero et al., 2010). Here we replicate the validation study across ZCTAs across three levels of urbanicities according to the Census (Federal Register). The correlation coefficients r = 0.995, 0.996, and 0.992 are highly consistent across urban ($n = 54$), low-density ($n = 121$), and rural areas ($n = 981$), respectively. A similar analysis on confusion matrices reveals that the precision ratios are 90.5%, 93.2%, and 92.9% across urban, low-density, and rural areas, respectively. That is to say, low-density and rural ZCTAs enjoy slightly higher precisions than urban ZCTAs in terms of whether the interpolated cancer incidences fall on the right sides of suppression threshold. In conclusion, our method generates the results that are largely consistent across the urban-rural spectrum and does not suffer from the vulnerability in rural areas.

To further validate our method, we compare the result to several methods commonly used for data interpolation. These methods include Ordinary Least Squares (OLS) regression, Generalized Linear Model (GLM), Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBoost). The predictors (explanatory variables) are the population counts for subgroups and the response variable is the cancer incidences with $n = 1156$ at the ZCTA level. They are the same as those used in the Monte Carlo simulation method.

Fig. 12 plots the results for the five methods and suggest that OLS and GLM do not yield as good results as the other three methods. Table 3 reports the two common measures of performance, MAE and r, which are used in assessing our Monte Carlo simulation method. Our method outperforms all five methods. For interpolation of suppressed cancer data, XGBoost may represent the best choice next to ours but still underperforms significantly. Similarly, we also compile the confusion matrix to assess the precision level for each method similar to the result on the Monte Carlo simulation method as reported in Table 2. The precision rate of our method is much higher than the other five methods.

For ML models,we also tested the inclusion of ancillary variables, such as poverty, health insurance, and education levels, in the models. However, these variables introduced noise and redundancy, leading to decreased model performance, as evidenced by increased MAE and RMSE values. Consequently, we focus on robust population and demographic data to optimize predictive accuracy. As shown in Supplementary Tables S1 and S2, XGB achieves the best performance when using population data alone demonstrating its superior accuracy in predicting case counts. GBM shows greater robustness after incorporating ancillary variables, but still underperforms when compared to XGB using only population data. Conversely, Random Forest and other models experience substantial increases in MAE and RMSE upon the inclusion of ancillary data, implying that such variables introduce more noise than signal, thereby degrades model performance. Moreover, when using only population data, most predictions remain positive. However, the inclusion of ancillary variables leads to an increased number of predictions below zero, indicating potential overfitting or instability introduced by these variables (Supplementary Fig. S1). This further supports the conclusion that ancillary data may not contribute meaningful predictive power in this specific context and, instead, could undermine model reliability.

5. Discussion

This paper aims to interpolate suppressed cancer data at fine spatial scales to maintain hierarchical total volume consistency and robust performance across rural and urban areas. The innovative multi-constraint Monte Carlo simulation method is effective in estimating suppressed county-level cancer counts and further disaggregating these counts to the ZIP Code Tabulation Area (ZCTA) level. The methodology leverages existing demographic structure as risk factors and reported cancer counts at higher geographic levels or in larger demographic groups as constraints and ensures a consistent and accurate imputation of cancer data across different geographic levels. The model comparison shows its advantages over traditional machine learning and regression methods for suppressed cancer data estimation.

The case study in Utah largely validates the effectiveness of the method. Our findings substantiate the robustness of population-weighted approaches (Curriero et al., 2010), especially for health data as reported in recent literature (Behal et al., 2023; Jones et al., 2020). By applying Monte Carlo simulations with simple constraints to improve dynamic proportion constraints during the case generation process, our study extends the capabilities of restricted and controlled Monte Carlo methods (RCMC) in the context of cancer data (Shi et al., 2013). Our method’s high accuracy across areas of various urbanicities alleviates the concern of less accuracy for rural cases by population-based interpolation methods (Dilekli et al., 2018). Our results also suggest that geo-imputation methods can achieve high effectiveness even without considering spatial correlation. The method provides a reliable framework for accurately estimating suppressed cancer cases in small population subgroups or geographic areas with small population and demonstrates its ability to manage uncertainty and improve prediction (Georgati et al., 2024; Scheiter et al., 2022). Among the machine learning models tested, XGB shows the best potential in handling complex spatial data imputation tasks (Wilson et al., 2022), but still underperforms the multi-constraint Monte Carlo simulation method.

The performance decline observed with ancillary data highlights the importance of careful variable selection in health data modeling. Variables that are weakly correlated with cancer incidence or are prone to data quality issues may add complexity without improving model accuracy. This finding aligns with existing studies (Amitha et al., 2021; Doshi et al., 2023), which emphasize that introducing redundant or noisy variables can negatively impact machine learning models. This confirms the applicability and accuracy of the Monte Carlo approach in scenarios with data suppression, particularly in complex contexts involving multidimensional population constraints (Hu et al., 2015; Huang et al., 2022).

6. Conclusion

This study demonstrates the effectiveness of the multi-constraint Monte Carlo method in estimating suppressed cancer data and downscaling to fine spatial scales. The proposed method achieves robust performance by maintaining hierarchical consistency and accurately addressing data suppression issues. However, while our results validate the method's utility in Utah, its generalizability to other regions and health conditions remains an avenue for future exploration.

First, extending the method to other geographic regions or disease types—such as chronic or infectious diseases—would allow for validation under diverse population structures and data suppression rules. Secondly, while this study does not explicitly account for spatial autocorrelation, incorporating spatial statistical techniques, such as geographically weighted regression or spatial Bayesian models, could enhance the accuracy of the estimates by addressing potential geographic clustering of cancer incidence. Thirdly, privacy concerns in health data remain a critical challenge. Future studies could integrate differential privacy or advanced anonymization techniques into the Monte Carlo framework to improve both data security and precision. Fourthly, moving beyond static five-year averages to incorporate dynamic time-series data would enable the capture of temporal trends, providing deeper insights into changing cancer incidence patterns. Finally, integrating Monte Carlo simulations with machine learning models offers significant potential for developing hybrid approaches that leverage the strengths of both paradigms to enhance prediction accuracy and applicability.

By addressing these areas, the proposed method will not only advance the state-of-the-art in suppressed health data imputation but also lay a foundation for broader applications in privacy-preserving public health data analysis. These directions underscore the flexibility and scalability of the Monte Carlo approach, reinforcing its potential for addressing critical challenges in health data analysis across varying contexts and constraints.

CRediT authorship contribution statement

Lingbo Liu: Writing – original draft, Visualization, Validation, Formal analysis, Data curation. **Lauren Cowan:** Resources, Data curation. **Fahui Wang:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization. **Tracy Onega:** Writing – review & editing, Resources, Project administration, Funding acquisition, Data curation.

Acknowledgement

The research is funded by the National Cancer Institute (NCI) under Grant No. R01CA267990-01 (Onega and Wang as MPIs) , and partially supported by National Science Foundation (NSF) grant #1841403. The 2016–2020 cancer data in Utah were provided by the Utah Cancer Registry (UCR). The UCR is funded by the NCI's SEER Program, Contract No. HHSN261201800016I, the U.S. Centers for Disease Control and Prevention's National Program of Cancer Registries, Cooperative Agreement No. NU58DP007131, with additional support from the University of Utah and Huntsman Cancer Institute. Points of view or opinions in this article are those of the authors and do not necessarily represent the official position or policies of the NCI or the UCR.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.healthplace.2024.103411.

Data availability

The data that has been used is confidential.

References

Amitha, P., Binu, V.S., Seena, B., 2021. Estimation of missing values in aggregate level spatial data. Clinical Epidemiology and Global Health 9, 304–309.

Avanasi, R., et al., 2016. Impacts of geocoding uncertainty on reconstructed PFOA exposures and their epidemiological association with preeclampsia. Environ. Res. 151, 505–512.

Baker, J., White, N., Mengersen, K., 2014. Missing in space: an evaluation of imputation methods for missing data in spatial analysis of risk factors for type II diabetes. Int. J. Health Geogr. 13, 47.

Behal, R., Davis, K., Doering, J., 2023. A novel adaptation of spatial interpolation methods to map health attitudes related to COVID-19. In: BMC Proceedings. Springer.

Bentley, G.C., Cromley, R.G., Atkinson-Palombo, C., 2013. The network interpolation of population for flow modeling using dasymetric mapping. Geogr. Anal. 45 (3), 307–323.

Boulton, A.J., Williford, A., 2018. Analyzing skewed continuous outcomes with many zeros: a tutorial for social work and youth prevention science researchers. J. Soc. Soc. Work. Res. 9 (4), 721–740.

Bozigar, M., et al., 2020. A geographic identifier assignment algorithm with Bayesian variable selection to identify neighborhood factors associated with emergency department visit disparities for asthma. Int. J. Health Geogr. 19, 1–16.

Buchin, K., et al., 2012. Processing aggregated data: the location of clusters in health data. GeoInformatica 16 (3), 497–521.

Buzaianu, E.M., Chen, P., Panchapakesan, S., 2017. Selecting the normal population with the smallest variance: a restricted subset selection rule. Commun. Stat. Theor. Methods 46 (16), 7887–7901.

Chang, E.T., et al., 2014. Validity of geographically modeled environmental exposure estimates. Crit. Rev. Toxicol. 44 (5), 450–466.

Chien, L.-C., Yu, H.-L., Schootman, M., 2013. Efficient mapping and geographic disparities in breast cancer mortality at the county-level by race and age in the US. Spatial and spatio-temporal epidemiology 5, 27–37.

Comber, A., Proctor, C., Anthony, S., 2008. The creation of a national agricultural land use dataset: combining pycnophylactic interpolation with dasymetric mapping techniques. Trans. GIS 12 (6), 775–791.

Cook, L.A., Sachs, J., Weiskopf, N.G., 2021. The quality of social determinants data in the electronic health record: a systematic review. J. Am. Med. Inf. Assoc. 29 (1), 187–196.

Cromley, R.G., Hanink, D.M., Bentley, G.C., 2012. A quantile regression approach to areal interpolation. Ann. Assoc. Am. Geogr. 102 (4), 763–777.

Curriero, F.C., et al., 2010. Using imputation to provide location information for nongeocoded addresses. PLoS One 5 (2), e8998.

Dilekli, N., et al., 2018. Evaluation of geoimputation strategies in a large case study. Int. J. Health Geogr. 17, 1–13.

Doshi, R., et al., 2023. Artificial intelligence's significance in diseases with malignant tumours. Mesopotamian Journal of Artificial Intelligence in Healthcare 2023, 35–39.

Federal Register:Urban Area Criteria for the 2020 Census-Final Criteria. 2022 03/24/2022 [cited 2024 10/10/2024]; Available from: https://www.federalregister.gov/documents/2022/03/24/2022-06180/urban-area-criteria-for-the-2020-census-final-criteria.

Fisher, P.F., Langford, M., 1996. Modeling sensitivity to accuracy in classified imagery: a study of areal interpolation by dasymetric mapping. Prof. Geogr. 48 (3), 299–309.

Georgati, M., et al., 2024. Modeling population distribution: a visual and quantitative analysis of gradient boosting and deep learning models for multi-output spatial disaggregation. Trans. GIS 28 (2), 130–153.

Goodchild, M.F., Lam, N.S.N., 1980. Areal interpolation - a variant of the traditional spatial problem. Geo Process. 1 (3), 297–312.

Gupta, S.S., Sobel, M., 1962. On selecting a subset containing the population with the smallest variance. Biometrika 49 (3–4), 495–507.

Harris, R., Chen, Z., 2005. Giving dimension to point locations: urban density profiling using population surface models. Comput. Environ. Urban Syst. 29 (2), 115–132.

Harvey, J.T., 2002. Population estimation models based on individual TM pixels. Photogramm. Eng. Rem. Sens. 68 (11), 1181–1192.

Henry, K.A., Boscoe, F.P., 2008. Estimating the accuracy of geographical imputation. Int. J. Health Geogr. 7, 3.

Howlader, N., et al., 2012. Use of imputed population-based cancer registry data as a method of accounting for missing information: application to estrogen receptor status for breast cancer. Am. J. Epidemiol. 176 (4), 347–356.

Hu, Y.J., Wang, F.H., 2015. Decomposing excess commuting: a Monte Carlo simulation approach. J. Transport Geogr. 44, 43–52.

Huang, L., et al., 2007. Detection of spatial clusters: application to cancer survival as a continuous outcome. Epidemiology 18 (1), 73–87.

Huang, C.F., Joseph, V.R., Mak, S., 2022. Population quasi-Monte Carlo. J. Comput. Graph Stat. 31 (3), 695–708.

Jones, R.R., et al., 2020. Impact of geo-imputation on epidemiologic associations in a study of outdoor air pollution and respiratory hospitalization. Spatial and Spatio-temporal Epidemiology 32, 100322.

Joseph Sheehan, T., et al., 2004. The geographic distribution of breast cancer incidence in Massachusetts 1988 to 1997, adjusted for covariates. Int. J. Health Geogr. 3, 1–12.

Kim, J.S., Gao, X., Rzhetsky, A., 2018. RIDDLE: race and ethnicity imputation from disease history with deep LEarning. PLoS Comput. Biol. 14 (4).

Kim, D., Chun, Y., Griffith, D.A., 2024. Impacts of spatial imputation on location-allocation problem solutions. Spatial Statistics 59, 100810.

Klassen, A.C., Kulldorff, M., Curriero, F., 2005. Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. Int. J. Health Geogr. 4, 1–16.

Kyriakidis, P.C., Yoo, E.H., 2005. Geostatistical prediction and simulation of point values from areal data. Geogr. Anal. 37 (2), 124–151.

Lam, N.S.N., 1983. Spatial interpolation methods - a review. Am. Cartogr. 10 (2), 129–149.

Li, J., Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. Ecol. Inf. 6 (3–4), 228–241.

Li, M., Baffour, B., Richardson, A., 2020. Bayesian spatial modelling of early childhood development in Australian regions. Int. J. Health Geogr. 19 (1), 43.

Lipscomb, J., et al., 1998. Predicting the cost of illness: a comparison of alternative models applied to stroke. Med. Decis. Making 18 (2 Suppl. l), S39–S56.

Liu, L., 2024. An ensemble framework for explainable geospatial machine learning models. Int. J. Appl. Earth Obs. Geoinf. 132, 104036.

Liu, X.H., Kyriakidis, P.C., Goodchild, M.F., 2008. Population-density estimation using regression and area-to-point residual kriging. Int. J. Geogr. Inf. Sci. 22 (4), 431–447.

Liu, L., et al., 2018. Exploring urban spatial feature with dasymetric mapping based on mobile phone data and LUR-2SFCAe method. Sustainability 10 (7), 2432.

Luo, L., McLafferty, S., Wang, F., 2010. Analyzing spatial aggregation error in statistical models of late-stage cancer risk: a Monte Carlo simulation approach. Int. J. Health Geogr. 9, 51.

Martin, D., 1989. Mapping population-data from zone centroid locations. Trans. Inst. Br. Geogr. 14 (1), 90–97.

Naumova, E.N., 2022. Precision public health: is it all about the data? J. Publ. Health Pol. 43 (4), 481.

Reibel, M., Agrawal, A., 2007. Areal interpolation of population counts using pre-classified land cover data. Popul. Res. Pol. Rev. 26 (5–6), 619–633.

Sahar, L., et al., 2019. GIScience and cancer: state of the art and trends for cancer surveillance and epidemiology. Cancer 125 (15), 2544–2560.

Scheiter, M., Valentine, A., Sambridge, M., 2022. Upscaling and downscaling Monte Carlo ensembles with generative models. Geophys. J. Int. 230 (2), 916–931.

Schroeder, J.P., Van Riper, D.C., 2013. Because Muncie's densities are not Manhattan's: using geographical weighting in the expectation–maximization algorithm for areal interpolation. Geographical analysis 45 (3), 216–237.

Shi, X., et al., 2013. Mapping disease at an approximated individual level using aggregate data: a case study of mapping New Hampshire birth defects. Int. J. Environ. Res. Publ. Health 10 (9), 4161–4174.

Taparra, K., Pellegrin, K., 2022. Data aggregation hides Pacific Islander health disparities. Lancet 400 (10345), 2–3.

Tobler, W.R., 1979. Smooth pycnophylactic interpolation for geographical regions. J. Am. Stat. Assoc. 74 (367), 519–530.

Walter, S.R., Rose, N., 2013. Random property allocation: a novel geographic imputation procedure based on a complete geocoded address file. Spatial and spatio-temporal epidemiology 6, 7–16.

Wan, H., et al., 2023. Areal interpolation of population projections consistent with different SSPs from 1-km resolution to block level based on USA Structures dataset. Comput. Environ. Urban Syst. 105, 102024.

Wang, F., et al., 2020. Automated delineation of cancer service areas in northeast region of the United States: a network optimization approach. Spatial and spatio-temporal epidemiology 33, 100338.

Wilson, T., et al., 2022. Methods for small area population forecasts: state-of-the-art and research needs. Popul Res Policy Rev 41 (3), 865–898.

Wright, J.K., 1936. A method of mapping densities of population with cape cod as an example. Geogr. Rev. 26 (1), 103–110.

Xie, Y.C., 1995. The overlaid network algorithms for areal interpolation problem. Comput. Environ. Urban Syst. 19 (4), 287–306.

Zhao, X., Curtis, A., 2024. Bayesian inversion, uncertainty analysis and interrogation using boosting variational inference. J. Geophys. Res. Solid Earth 129 (1) e2023JB027789.

Zhu, D., et al., 2020. Spatial interpolation using conditional generative adversarial neural networks. Int. J. Geogr. Inf. Sci. 34 (4), 735–758.