Empirically-Derived Synthetic Populations to Overcome Small Sample Sizes

Erin E. Fowler[1], Anders Berglund[2], Thomas A. Sellers[1], Steven Eschrich[2], John Heine[1,*]

Moffitt Cancer Center & Research Institute, 12901 Bruce B. Downs Blvd, Tampa FL, 33612 (MCC)

[1]Cancer Epidemiology Department, MCC

[2]Department of Biostatistics and Bioinformatics, MCC

[*]Corresponding Author

Email Addresses
EEF: erin.fowler@moffitt.org
AB: anders.berglund@moffitt.org
TAS: thomas.sellers@moffitt.org
SE: steven.eschrich@moffitt.org
JH: john.heine@moffitt.org

Abstract ~ Overfitting is a problem often encountered when developing multivariate predictive models with limited data. The objective of this work is to present a method to generate a synthetic population (SP) from a sparse seed dataset that has a similar multivariate structure to aid in model building. We used a multivariate kernel density estimation approach with an unconstrained bandwidth to generate SPs with data at the individual level. A matched case-control study (n =180 pairs) was used as the seed dataset. Cases and controls were considered as two subpopulations and analyzed separately. We included four continuous measures and one categorical variable for each subject. Bandwidth matrices were determined with differential evolution (DE) optimization based on covariance comparisons. Similarity between the seed dataset with datasets selected randomly from SPs were compared under the hypothesis that the structure should be similar. To evaluate similarity, we compared PCA score distributions and residuals summarized with the distance to the model in X-space (DModX).

SPs were generated for both case and control groups. The probability of selecting seed replicas when constructing synthetic sample datasets randomly was minute. Within group, both PCA scores and residuals were similar across seed and synthetic samples; covariance comparisons also indicated the structure was similar.

Feasibility of a new SP generation methodology was presented. This approach produced synthetic data at the patient level indistinguishable from the seed data. The methodology coupled kernel density estimation with DE optimization and deployed novel similarity metrics derived from PCA. The use of synthetic samples may be useful to mitigate overfitting in initial model building exercises. To further develop this approach into a research tool for model building, additional evaluation with increased dimensionality is required.

Keywords: Synthetic data generation, kernel density estimation, overfitting, differential evolution, principal component analysis, DModX

1. Background

Biomedical research often generates high-dimensionality datasets. High-dimensionality can pose significant challenges when the goal is to translate the research data into multivariate predictive models that replicate beyond initial testing. Initial findings can lack rigor and reproducibility due to overfitting. In this situation, a fitted model describes random error or idiosyncrasies of the training data rather than the intended process (1), which is a common artifact to many areas of bioinformatics (2). Overfitting is typically encountered when there are many predictor variables and model parameters relative to the sample size during the developmental stages, resulting in subsequent model failure even with lower dimensionality as well.

Established guidelines (2) for high-dimensionality datasets define two stages for developing predictive tests for clinical applications: stage 1 includes discovery and test validation phases; and stage 2 includes clinical utility evaluation. The first stage includes full model specification and independent confirmation. For predictive models, all parameters are fixed in stage 1. In the later stage, utility is tested in a clinical trial. These guidelines apply under more general conditions as well. Harrell (3) proposed a rough rule of thumb to decrease the risk of overfitting when modeling a binary outcome. Briefly, the number of variables, or more aptly degrees of freedom, considered when model building should be within this range m/20 - m/10 range, where m is the number of samples of the smaller group, and it includes all variables examined even if they did not make it into the specified model (4). Automated variable selection, pretesting potential predictor variables, and dichotomizing continuous variables are all contributing factors to spurious findings (1). The extent that triaged variables and analyses are discussed in the literature is limited. It may not always be possible to collect sufficient data during the initial development phase due to time constraints on patient accrual, or cost, to test potential models thoroughly. It is our premise that it is better to make a predictive model fail initially rather than with more costly studies in the future. Methods that can be employed to mitigate overfitting can contribute importantly in situations when limited sample sizes exist.

There are a number techniques to address overfitting (5) when sample sizes do not permit independent evaluation. These include splitting the dataset into training and independent (validation) subsets and various forms of resampling such as cross-validation and bootstrapping. Splitting the sample provides for independent replication but is dependent upon a dataset that is large enough to separate. Leave-one-out cross-validation produces a stringent test that underestimates the performance but has a drawback in that some instances removing one sample from the dataset will not offer enough variation (5). More generally, cross-validation can be highly variable (6). Bootstrapping (6) is an alternative to cross-validation but does not offer an independent validation set. Cross-validation depends on repeated re-training to estimate the error rates. As such, it does not produce a fitted model as an end product. As a potential remedy, we will borrow from synthetic population generation methods (7) and explore generating synthetic patients with the relevant variables for model development purposes.

Synthetic population (SP) generation methods are frequently used in policy planning studies (8), such as land use and transportation research (9-11) when accessing data may be prohibitive due to privacy or cost issues (9). The objective is to construct joint probability for the population of interest (target) given the aggregated data (marginal distributions or tabulated data) and a sample of disaggregated data (joint distribution) with specified constraints. Established approaches (7, 8, 12) include deterministic re-weighting techniques and methods that use some form of stochastic component in the synthesis. Iterative proportional fitting (IPF) is deterministic re-weighting method that is a common approach (13) based on iteratively adjusting contingency tables relative to the given constraints as detailed by Lomax and Norman (14). Although IPF extends to any number of dimensions (15), there are noted efficiency problems for large dimensionality (16). The limitations with IPF are discussed in detail elsewhere (11, 14, 16); concerns relative to our work are that too many zero cells (cells with zero value) hinder convergence and high-dimensionality. The conditional probability approach uses the sampled distributions to generate the target population by randomly sampling one variable at a time but becomes cumbersome for a larger number of variables, and the outcome is dependent upon how the constraints are introduced (i.e. ordered) into the stochastic reconstruction (12). Combinatorial optimization techniques

(CO) are stochastic methods that include the hill-climbing and simulated annealing (12, 17). In the CO approach, the sample population (seed) is often assumed large compared to a subpopulation of interest (or target population). Random samples are drawn from the larger population and compared with the target population constraints; the element that enhances the goodness of fit becomes part of the target population (12, 17, 18). Hill-climbing has the advantage of speed (17) but can become trapped in suboptimal solutions (12, 17). Simulated annealing is a refinement of hill-climbing that permits the possibility of backing out of a suboptimal solution but as a consequence may lose performance (12, 17).

In this report, we present a new approach for generating SPs and present initial findings. Our boundary conditions are different than some of the approaches discussed above. We start with a sparse seed sample of the aggregated data with the goal of building a larger dataset that resembles this sample. We use multivariate kernel density estimation with an unconstrained bandwidth matrix to estimate the underling probability density function (pdf) initiated with a sparsely populated seed population. The objective is to generate synthetic populations with individual level data. We use Differential Evolution (DE) optimization (19) to determine the bandwidth parameters for the kernel density estimation by matching the covariance structure of the seed population. Novel similarity metrics based on principal component analysis (PCA) were employed to compare the covariance structure between the seed population and random samples drawn from the SP. When fully developed, this technique could produce synthetic samples for model building exercises, providing the underlying data structure of the seed sample was captured in the synthetic construction. Once a model is optimized, it can be initially validated with the seed population producing a fully specified model suitable for independent validation. In this report, we outlay our methodology and demonstrate the feasibility of our approach by synthesizing a matched case-control dataset. To put our task in context with the synthetic techniques discussed above, our seed population is a sparse sample of the target joint density.

2. Methods and Materials

2.1 Seed and synthetic populations

A matched case control study (n = 180 pairs) comprised of women that underwent mammography at this Center was used as the seed dataset. Women with first time unilateral breast cancer define the case population. The control population included women without a history of breast cancer. Controls were individually matched to breast cancer cases by age (±2 years), and other factors not relevant to this work. The data accrual and nuances of the matching were discussed previously (20-23). Although this dataset was constructed and investigated for epidemiologic purposes, we use it to present our SP generation methodology in a more realistic situation, where the epidemiologic context is irrelevant. For feasibility, we selected these variables for each patient: (1) age measured in years [yr]; (2) height measured in inches [in], (3) mass [kg], (4) breast density measure (percentage) , referred to as PD; and (5) menopausal status, labeled as MS. Measures 1-4 are continuous variables rounded to integer values and measure 5 is binary with MS = 0 for menopausal women and MS = 1 otherwise. We note, we used inches for height because larger metrics (meters or feet) were too coarse, whereas a smaller metric, such as cm, was too fine.

Synthetic data was generated similar to the way the seed dataset was established. Because controls were matched to cases (not selected randomly), we generated two synthetic populations. For the synthetic case population, $s^1(\mathbf{x})$, we considered four integer variables referred to as the vector **x**, and generated the conditional population given MS =1, $s^1(\mathbf{x}\,|\mathrm{MS}=1)$, and conditional population given MS = 0, $s^1(\mathbf{x}\,|\mathrm{MS}=0)$. The synthetic case population was constructed by adding these conditional populations with the respective weights

$$s^1(\mathbf{x}) = s^1(\mathbf{x}\,|\mathrm{MS}=1) \times p_{11} + s^1(\mathbf{x}\,|\mathrm{MS}=0) \times p_{10}\,, \qquad \text{Eq.[1]}$$

where p_{11} and p_{10} are the respective weights determined from the sample; $p_{11} = 0.773$ and $p_{10} = 1.0 - p_{11}$. A realization of a case was produced by taking a random sample from the case population using the methods discussed below. The control population was constructed similarly using the respective conditional populations expressed as

$$s^0(\mathbf{x}) = s^0(\mathbf{x}\,|\mathrm{MS}=1) \times p_{01} + s^0(\mathbf{x}\,|\mathrm{MS}=0) \times p_{00}\,, \qquad \text{Eq.[2]}$$

where p_{01} and p_{00} are the respective weights for the controls: $p_{01} = 0.788$ and $p_{00} = 1.0 - p_{01}$. In a modeling situation, a given control realization must be matched to its case. This can be achieved by drawing a random sample from a restricted region of s^0 corresponding to the matching variables. In the analyses, the populations were evaluated separately forgoing the matching necessity. Thus, we are treating the case and control seed populations as two separate entities in this work. This will provide two varied populations to evaluate our methods.

2.2 Population Density Estimation

To generate synthetic data, we used a multivariate kernel density approach to estimate the underlying joint density of the target population given the seed population. We define an m component row vector for the ith seed sample (one sample from the parent population) $\mathbf{x}_i = (x_1, x_2, x_3 \ldots x_m) = [x_1, x_2, x_3 \ldots x_m]^T$ from n samples (i.e. patients); in this example n = 180 and m = 4. The components of $\mathbf{x}_i$ are x_{ij} and are sometimes referred to as features or predictor variables. In this work, we rounded all variables to integer values to reduce the dimensionality of the density estimation. We define the prospective m component column vector as $\mathbf{x}$ with components as x_j, which is a synthetic entity. The parent population assumes the role of the training dataset for the SP generation. For motivational purposes, we first illustrate the simpler formulism. It may be shown that the estimated joint density function, $s_0(\mathbf{x})$, for a sampled population is given by (24) for the constrained estimation

$$s_0^k(\mathbf{x}) = \frac{1}{c_x\, n_k} \sum_{i=1}^{n} \exp\left[-\sum_{j=1}^{m} \frac{(x_j - x_{ij})^2}{2\sigma_j^2} \right], \qquad \text{Eq.[3]}$$

where we have assumed independent samples, c_x is a normalization constant, σ_j are the bandwidth parameters, n_k is the number of samples in the seed population, and we are using a Gaussian kernel noting there are many valid kernels. The super/subscript k indicates case (k = 1) and control (k = 0), noting the two subpopulations are generated separately. This operation is performed for each $\mathbf{x}$. Equation [3] is too restrictive for our application because it lacks variable interaction. The above expression can be generalized to capture the relevant covariance structure (25)

$$s_0^k(\mathbf{x}) = \frac{1}{c_x\, n_k}\sum_{i=1}^{n} \exp\left(-\frac{1}{2}[\mathbf{x}-\mathbf{x}_i]^T\mathbf{H}^{-1}[\mathbf{x}-\mathbf{x}_i]\right) = \frac{1}{n}\sum_{i=1}^{n} k_H(\mathbf{x}), \qquad \text{Eq.[4]}$$

giving the unconstrained bandwidth expression. **H** is the bandwidth matrix (here, related to the covariance), which is symmetric and positive definite. These conditions on **H** ensure each term in the sum provides a valid probability density increment. If **H** is diagonal, the above expression reduces to Eq.[3]. Initial experimentation showed we could drop the ½ scale factor in the argument of Eq.[4] and estimate (initialize) **H** as the covariance.

The bandwidth parameters affect both the degree of smoothing and orientation. Selecting the optimal bandwidth parameters for unconstrained **H** is critical and non-trivial. In particular, bandwidth selection for the unconstrained multivariate is the subject on ongoing research, (26, 27). We use DE optimization based on covariance comparisons. We define the n × m design matrix for the seed population as $\mathbf{X}_m$ with elements $\mathbf{x}_{ij,}$ where the rows of $\mathbf{X}_m$ contain $\mathbf{x}_i^T$. We mean center the columns of $\mathbf{X}_m$ giving **X** and form the covariance for the seed population given by

$$C = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}, \qquad \text{Eq.[5]}$$

where **C** is an n×m matrix. We define the covariance of a random sample drawn from a SP as $\mathbf{C}^s$ using the same formulism expressed in Eq.[5]. For the optimization procedure, we used the normalized absolute difference matrix **E**, with elements $e_{ij} = \left|\frac{c_{ij}-c_{ij}^s}{c_{ij}}\right| \times 100$. **H** is specified by minimizing the difference

$$\varepsilon = \sum_{ij} e_{ij}\,. \qquad \text{Eq.[6]}$$

We provide a brief description of DE aided by the schema shown in Figure 1 and Figure 2. Where possible, we follow the notation established by its founders (19). This is a global optimization technique that constructs a

solution via competition. We start with a randomly initialized vector field (zeroth generation) where the components, indexed by j, of a given vector correspond to the elements of **H**. We used a uniform random variable to generate these components constrained to ±0.10×estimated respective covariance from the seed data. This population is comprised of Np vectors (i.e. the population size) with D_n components, where g is the generation index. We refer to the entire population as $\mathbf{w}_g$ and label its members as $\mathbf{w}_{ig}$, where i is the member index. For this development, D_n is the number of unique elements in **H**. For a given g, we cycle through i = 1 to Np (Np is defined below) and select $\mathbf{w}_{ig}$ at each increment illustrated in Figure 1. A mutant vector is constructed by adding a weighted difference between two vectors to third vector all chosen at random from $\mathbf{w}_g$ expressed as

$$\mathbf{v}_{ig} = \mathbf{w}_{qg} + F \times \left(\mathbf{w}_{rg} - \mathbf{w}_{sg}\right), \qquad \text{Eq.[7]}$$

subject to this condition: $q \neq r \neq s \neq i$, which induces diversity. The weight, F, is referred to as the amplification factor, which controls the mutation size or effect. The mutant vector is used to construct a trial vector defined as $\mathbf{u}_{ig}$ that will compete with $\mathbf{w}_{ig}$ shown in Figure 1. A trial vector is constructed by permitting the possibility (at random) of crossing components of this mutant vector with the components of the target vector dictated by a crossover rate (CR) shown in Figure 2. The crossover operation occurs at the vector component level and dictates the mutation rate. This is to enhance the possibility that the trial vector has at least one component from the mutation (Figure 2). There are two avenues for the mutation to occur at each component: if CR < a uniform [0,1] rv or if a random [1,n] integer = j. The random quantities are defined as c_j and d_j, respectively. A given $\mathbf{u}_{gi}$ represents a mutation (most of the time) of $\mathbf{w}_{ig}$ derived from Eq.[7]. A given set of target and trial vectors compete to move the next generation by comparing their respective ε calculated with Eq.[6] by generating an SP for each vector. The vector that produces the smaller ε moves to the next to generation; Np comparisons of $\mathbf{w}_{ig}$ with $\mathbf{u}_{ig}$ produce the next generation population defined as $\mathbf{w}_{g+1}$. Thus, successive generations of $\mathbf{w}_g$ become incrementally more fit. This process is performed generation by generation (Figure 1) until convergence. Although **H** has m^2 components, we force symmetry on the solution and use m×(m+1)/2 components; thus all the vectors in the DE application have D_n = m×(m+1)/2 components. For the population

size we use Np = 10×D_n= 10×10=100. We also force the positive definite property on the solution by checking **H** for positive Eigenvalues. When a possible solution is not positive definite we assign a large error term to the competition rather than using Eq.[6]. In the event that both vectors produce a non-positive definite **H**, the target vector moves to the next generation. Thus, as the generations unfold, the possibility of forming a non-positive definite potential solution tends to zero.

Figure 1: Differential Evolution Trial Competition: This illustrates the basic competition for one trial within a given generation. There are Np trial competitions per generation. The ith vector from the current generation, $\mathbf{w}_{ig}$, competes with a trial vector, $\mathbf{u}_{ig}$. The trial vector is a mutation of $\mathbf{w}_{ig}$ with attributes derived from the population $\mathbf{w}_g$ shown in Figure 2.

Figure 2: Trial Vector Construction: The trial vector, $\mathbf{u}_{ig}$ is constructed with components from $\mathbf{w}_{ig}$ and the mutant vector $\mathbf{v}_{ig}$. The $\mathbf{v}_{ig}$ construction is shown in Eq.[7]. The Vectors $\mathbf{u}_{ig}$ and $\mathbf{w}_{ig}$ compete shown in Figure 1.

2.3 Synthetic Population Construction

A sample population from the SP is constructed after specifying $s_0(\mathbf{x})$ with the optimization procedure. We adapted the standard method of mapping a uniformly distributed random variable to a random variable with a specified (target) probability distribution (pdf), sometimes referred to as inverse sampling. For completeness, we outline this standard transform method to put our approach in context. We define two random variables: u, defined over this range [0,1] with a pdf given by $p_u(u) = 1$, i.e. u is uniformly distributed random variable; and t over this range (- ∞,∞) with the target pdf labeled as $p_t(t)$, where p_t is arbitrary. We then determine the function q(u) = t, such that t is distributed as $p_t(t)$. The solution conserves probability using the cumulative functions

$$\int_0^u du = \int_{-\infty}^t p_t(z)dz \rightarrow \quad u = P_t\,(t)\,, \qquad \text{Eq.[8]}$$

where P_t is the cumulative function for t. In passing the above shows that mapping any random variable with its own cumulative function produces a uniform random variable. Although we use Eq.[8] for our solution, we show that standard and equivalent result expressed as

$$t = q(u) = P_t^{-1}(u). \qquad \text{Eq.[9]}$$

To construct the SP, we conditioned $s_0(\mathbf{x})$ with reasonable approximations. If we assume there are on the order W women in this population (W ~million), we normalize $s_0(x)$ giving $s_1 = W\times s_0(\mathbf{x})$. For all **x** meeting this condition $s_1(\mathbf{x}) < 1$, we set $s_1(\mathbf{x}) = 0$ giving $s(\mathbf{x})$, which defines the SP. This conditioning permits only whole synthetic samples. To sample the SP, we let u_1 be uniformly distributed integer valued random variable over this range [1, w], where w is the *area* under $s_1(\mathbf{x})$ and is equivalent to the number of women in the SP after the conditioning. We assemble all of the unique combinations of the coordinates $\mathbf{x}= (x_1,...,x_m)$ into a one dimensional variable, z_i ,where z_1 is combination 1, z_2 is combination 2, z_i is the ith combination, and continue the ordering to last combination labeled as z_k, where k is the number of unique combinations. We order the corresponding number of samples with each combination into another one dimensional variable, f_i , also with k elements. The number of unique combinations is given by $k = r_1 \times r_2\times\ldots\times r_m$, where r_l is the range of each variable (i.e. the distance between the minimum and maximum values noting that these are integer valued metrics). It is important to recognize that combination ordering and labels are not important as long as the correspondence between z_i and f_i is maintained, which can be expressed as $f_i=f_i(z_i)$. The cumulative function for f_i is given by

$$F_j = \sum_{i=1}^{j} f_i \qquad \text{Eq.[10]}$$

for j =1 to k (i.e., $F_k = w$). This is the discrete and scaled analog of P_t expressed in Eq.[8]. A given sample from the SP is found by generating a realization of the uniform random variable, u_1 defined above, and determining $F_{j-1} < u_1 \leq F_j$. The interval F_j-F_{j-1} gives f_j noting that f_j points to the **x** combination indexed by z_j producing one synthetic sample from the SP. We perform this operation repeatedly to generate a sample

population from the SP. This solution is analogous to solving for a particular value of t (equating with **x**) given one realization of u expressed in Eq.[9].

2.4 Evaluation Methods

The analytic endpoint is to ensure the covariance structure between the seed and synthetic samples are similar. In this situation we used the seed dataset as the reference and examined the similarity between synthetic data with the reference. We applied two unique forms of analysis based on principal component analyses (PCA). First, the multivariate sample distributions drawn from the synthetic population will be compared to the seed dataset using PCA. A PCA model derived from the seed population will be applied to the synthetic data, and predicted scores (expansion coefficients) will be compared. This will be accomplished by comparing the distribution of the original samples and the synthetic samples along each individual principal component. Strong outliers will be observable in the score plots, whereas the detection of moderate outliers requires another approach. In the second approach, the distance to the model in X-space (DModX) statistics will also be used to access differences between these populations, where the residuals are evaluated. In short, the residuals for the two populations should also be the same. We note the PCA expansion and residual are also orthogonal complements. The PCA model was calculated (trained) on the seed population and the variables were expanded with square and interaction-terms of the original variables. Each variable was centered and scaled to unit variance. The derived PCA model was then applied to the synthetic datasets to predict the scores and generate DModX quantities. P Residuals between the seed and synthetic data were compared with a t-test. CA models were calculated using Evince (version 2.7.9, Prediktera AB, Umea, Sweden). Box Plots with Violin lines were generated in MATLAB (R2017b, MathWorks Inc. Natick MA, USA) using the GRAMM toolbox (28).

3. Results

The goal of the synthetic population generation is to start with a sparse sample and reconstruct a larger population from which synthetic samples can be drawn. The data sparsity is illustrated in Figure 3 using the

control seed sample (n=180). The top left figure shows there are two women with age = 58yr and height = 64in in the seed data. Moving down the column of Figure 3 panes, we included expanded age ranges from 58-59, and 58-60. Moving right across the row of Figure 3 panes, we included expanded height ranges from 64-66in and 64-68in. The most populated pane (age 58-60, height 64-68in) is on the bottom right with 15 samples, or 8% of the seed sample.

The synthetic population was created using the approach described above. Reconstruction properties of this approach are illustrated in Figure 4. This figure shows a two-dimensional slice through the four dimensional reconstructed SP (left pane) at age = 58yr and height = 64in corresponding with the most sparse pane in Figure 3 (top left). Note the dense nature of this conditional distribution in comparison with Figure 3. Taking a profile along the horizontal direction at mass = 64kg gives the associated conditional density (not normalized) for the PD (breast density metric) shown in the right pane (before thresholding). The smooth continuous curve illustrates the properties of the kernel reconstruction.

Figure 4: Synthetic data reconstruction illustration. This shows a two-dimension slice through the synthetic distribution (left pane) for woman with age = 58 and height =64in. Mass (kg) is on the vertical axis and PD on the horizontal axis. This corresponds to the sparsest sample (n=2) in Figure 3. The dashed line marks a profile for a woman with mass = 64kg. The corresponding distribution for PD (breast density measure) is shown on the right before integer thresholding and normalization were applied.

One consideration when generating synthetic sample distributions is the degree to which the seed samples influence subsequent synthetic sampling. Similarity in synthetic samples can be over-estimated depending on the frequency of seed samples that are observed. We generated 1000 synthetic sample populations (n = 180) in succession using the control group. After each sample population generation, we checked to see if at least one replica from the seed data was included in the synthetic sample. The average number of synthetic population sample dataset constructions was eighteen (average) before a seed replica was included; 97.2% of the time there was only 1 seed sample responsible for the count to trigger and in 2.8% either 2-3 samples were responsible.

Characteristics of the synthetic sample population were evaluated using PCA comparisons with the seed data. The associated findings are shown in Figure 5. Figure 5A (top left) shows the PCA scores for PC1 and PC2 (first two principal components) for the case seed data. The first principal component explains 26% of the variation with the second explaining 18.2%, giving 44.2 % in total for the parent data (derived from the seed data).These principal components are plotted together with the predictions of the four different synthetic datasets; these predictions were based on the trained PCA models for the cases and control groups (two separate models trained with the seed data). We note, none of the SP sample populations contained seed samples. The score plots show a distinct separation in all datasets due to MS (menopausal status). Approximately 99% of the samples in the right group in Figure 3A had MS =1 (about 1% had MS =0), while the compact group to the left consisted of samples with MS = 0 (except one sample or approximately 0.14%). Interestingly, this artifact was a characteristic of the seed data that is also part synthetic data structure. The PCA score plot in Figure 3C shows similar behavior in that there are few differences between the seed and the synthetic datasets. Note, the separation due to MS was also captured. The second dimension (y-axis) is in the opposite direction for the control group compared to the case group. This is a common artifact in PCA because the signs are arbitrary. This is usually accounted for by multiplication of -1 (i.e. PCA is invariant under a sign change).

Figure 5: PCA models for seed and synthetic data. The first two principal components derived from the seed data for the Case group are shown in **A** together with the predictions of the synthetic datasets. There is no visible difference between the seed data and the four different synthetic datasets. The residuals, DModX, are shown in **B** as a boxplot with violin density lines. The differences in the residuals between the seed and synthetic data were not significant. Similar results are seen for the Control group in pane **C** and **D**.

Figure 5A illustrates that there is no visible difference between the seed population and the four different synthetic datasets generated, as represented by the PCA model elements PC1 and PC2. The variation not explained by the PCA model is shown in the DModX plot in Figure 5B for the case group. The DModX for each dataset is represented by a boxplot and the distributions are shown as violin plots. In each plot, the p-value

from the comparison with the seed residual is provided. The difference in each comparison was not significant. Because the DModX values across the 5 populations were statistically similar, the seed and synthetic sets were indistinguishable. The same analysis was also performed on the control group shown in Figure 5D, which produced similar findings

The covariance matrices for the control seed data and a synthetic population sample are shown in Table 1 and for the cases in Table 2. Synthetic quantities are similar to the respective seed quantities when considering intra-group pairwise comparisons. The exception is the age-mass pair. For this pair, the covariance confidence intervals span zero indicating zero correlation in the seed dataset. Random synthetic population samples shows this occurs approximately 50% of the time (both cases and controls) there are shifts above and below zero (not shown). For reference, the four bandwidth matrices used to generate the SPs are provided in Table 3.

Table 1. Covariance Matrices for the control group

Controls Seed n = 180				
	height (in)	**age**	**PD**	**mass (kg)**
height (in)	6.3 (5.1, 7.4)	-3.9 (-7.9, 0.04)	6.6 (1.6, 11.6)	3.9 (0.20, 7.7)
age	-3.9 (-7.9, 0.0)	108 (89, 128)	-41.4 (-65.4, -17.4)	-2.8 (-16.6, 10.9)
PD	6.6 (1.6, 11.6)	-41.4 (-65.4, -17.4)	213 (167, 258)	-53 (-72, -32)
mass (kg)	3.9 (0.21, 7.6)	-2.82 (-16.6, 10.9)	-52 (-73, -31)	116 (84, 148)

Controls Synthetic n = 180				
	height (inch)	**age**	**PD**	**mass (kg)**
height (inch)	6.5 (5.3, 7.8)	-3.92 (-7.4, -0.4)	11.4 (5.4, 17)	6.3 (2.5, 10)
Age	-3.9 (-7.4, -0.4)	106 (86, 125)	-411 (-62, -19)	-3.4 (-17, 10)
PD	11.4 (5.4, 17.4)	-41 (-62, -19)	185 (140, 230)	-36 (-57, -15)
mass (kg)	6.3 (2.5, 10.0)	-3.4 (-17, 10)	-36 (-57, -15)	128 (102, 154)

These tables show the covariance matric for the seed population (left) and one synthetic sample (right) for controls. Confidence intervals are provided parenthetically.

Table 2. Covariance Matrices for the case group

Cases n = 180				
	height (in)	**age (yr)**	**PD**	**mass (kg)**
height (in)	7.1 (5.6, 8.5)	-7.4 (-11.8, -3.0)	-1.64 (-8.3, 5.0)	9.6 (4.8, 14.4)
age	-7.4 (-11.8, -3.0)	109 (92, 128)	-43 (-69, -18)	-2.9 (-21.4, 15.6)
PD	-1.6 (-8.3, 5.0)	-43.3 (-69, -18)	246 (198, 293)	-81 (-109, -53)
mass (kg)	9.6 (4.8, 14.4)	-2.9 (-21.4, 15.6)	-81 (-109, -53)	166 (122, 209)

Cases Synthetic n = 180				
	height (in)	**age (yr)**	**PD**	**mass (kg)**
height (inch)	9.0 (7.2, 11)	-10 (-15, -5.4)	-0.3 (-8.5, 7.8)	15 (9.4, 20)
age	-10 (-15, -5.4)	117 (100, 133)	-47 (-72, -21)	-1.1 (-19, 17)
PD	-0.3 (-8.5, 7.8)	-47 (-72, -21)	220 (172, 269)	-55 (-81, -29)
mass (kg)	15 (9.4, 20)	-1.1 (-19, 17)	-55 (-81, -29)	140 (109, 172)

These tables show the covariance matric for the seed population (left) and one synthetic sample (right) for cases. Confidence intervals are provided parenthetically.

Table 3. Bandwidth Matrices

Controls Menopausal n = 132				
	Height (inch)	**Age**	**PD**	**Weight (kg)**
Height (inch)	5.16	-2.08	2.93	5.20
Age	-2.08	62.0	-14.4	-5.33
PD	2.93	-14.4	185.6	-48.1
Weight (kg)	5.20	-5.33	-48.1	141.3

Cases Menopausal n = 142				
	Height (inch)	**Age**	**PD**	**Weight (kg)**
Height (inch)	6.27	-4.94	-2.28	10.6
Age	-4.94	66.9	-3.69	-10.1
PD	-2.28	-3.69	188.3	-59.4
Weight (kg)	10.6	-10.1	-59.4	164.8

Controls non-Menopausal n = 48				
	Height (inch)	**Age**	**PD**	**Weight (kg)**
Height (inch)	6.68	2.14	11.9	2.23
Age	2.14	31.2	-24.8	-6.17
PD	11.9	-24.8	242.0	-65.3
Weight (kg)	2.23	-6.17	-65.3	73.3

Cases non-Menopausal n = 38				
	Height (inch)	**Age**	**PD**	**Weight (kg)**
Height (inch)	11	-0.41	-10.8	9.02
Age	-0.41	19.7	-10.4	-2.85
PD	-10.8	-10.4	314.8	-130.5
Weight (kg)	9.02	-2.85	-130.5	199.3

The bandwidth matrix, H, generated be DE is provided for each subgroup. These were used to generate the synthetic populations.

4. Discussion

We presented a novel methodology to generate a SP given a sparsely populated seed dataset. The approach is based on an unconstrained multivariate kernel density estimation driven by an optimization procedure. In this approach, we compared the covariance structure rather than considering the shape of the density relative to the seed points; often the performance of a kernel estimator is based on some form of squared error analysis taken globally (25, 29). The feasibility of our approach is evident in the similarity of the comparisons with PCA. A benefit of the PCA approach is that all variables under consideration are evaluated simultaneously and repeatedly because each PCA dimension is a linear combination of the input variables. For either cases or controls, the respective synthetic dataset samples exhibited similar behavior under the PCA transform. We presented this approach to generate SP samples for model building purposes. The approach can also be used as a method of imputation for missing variables noting the entire dataset would be used in the estimation. The method also lends itself to generating sampling requirements for multivariate investigations given a population sample.

Although our study addressed many technical challenges, there are several limitations and nuances worth noting. We considered five variables. This dimensionality was sufficient to develop a general framework and provide initial feasibility. To make the approach into a model building tool, the dimensionality of the input space must be increased necessitating increased data processing capabilities. Moreover, comparisons with accepted techniques such as cross validation and bootstrapping will be required to adequately rank this approach. We considered four continuous and one dichotomous variable. Often the input space could include a wider mix of continuous, ordinal, and nominal variables. We used a conditional probability approach to address the limited mixture of variables in this work. Additional methods of estimating the covariance between mixed variable types (30) are required to render the problem suitable for kernel processing and to reduce the number of conditional probability branches. Otherwise, extensive branching becomes a limitation for datasets with few samples. The process was driven by considering the covariance directly. It is not clear at this time, if this is generally optimal in wider settings. Alternatively, the optimization could be driven by these PCA similarity metrics or other summary metrics. Additional experimentation is required to determine the optimal method of driving the optimization process.

5. Conclusions

The work presented in this report contributes to both kernel density estimation and synthetic population research. The optimization feedback loop can be modified to incorporate other endpoints. The work also introduced two related methods of analyzing structure similarity with PCA. These similarity metrics have applications well beyond those presented in this report. The use of synthetic samples may be useful to mitigate overfitting in initial model building exercises providing our techniques is evaluated under more general conditions. Planned research includes addressing the limitations addressed above.

List of Abbreviations

Differential evolution (DE), the distance to the model in X-space (DModX), menopausal status (MS), principal component analysis (PCA), probability density function (pdf), random variable(rv), synthetic population (SP)

Declarations

Ethics and consent to participate: This study was approved by the Institutional Review Board (IRB), University of South Florida, Tampa, FL (IRB# Ame13_104715)

Consent for publication: The work does not contain personal identifiers.

Availability of data: The datasets used and/or analyzed 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 work was supported by Moffitt Cancer Center grant #17032001 (Miles for Moffitt) and National Institutes of Health Grant #R01CA114491.

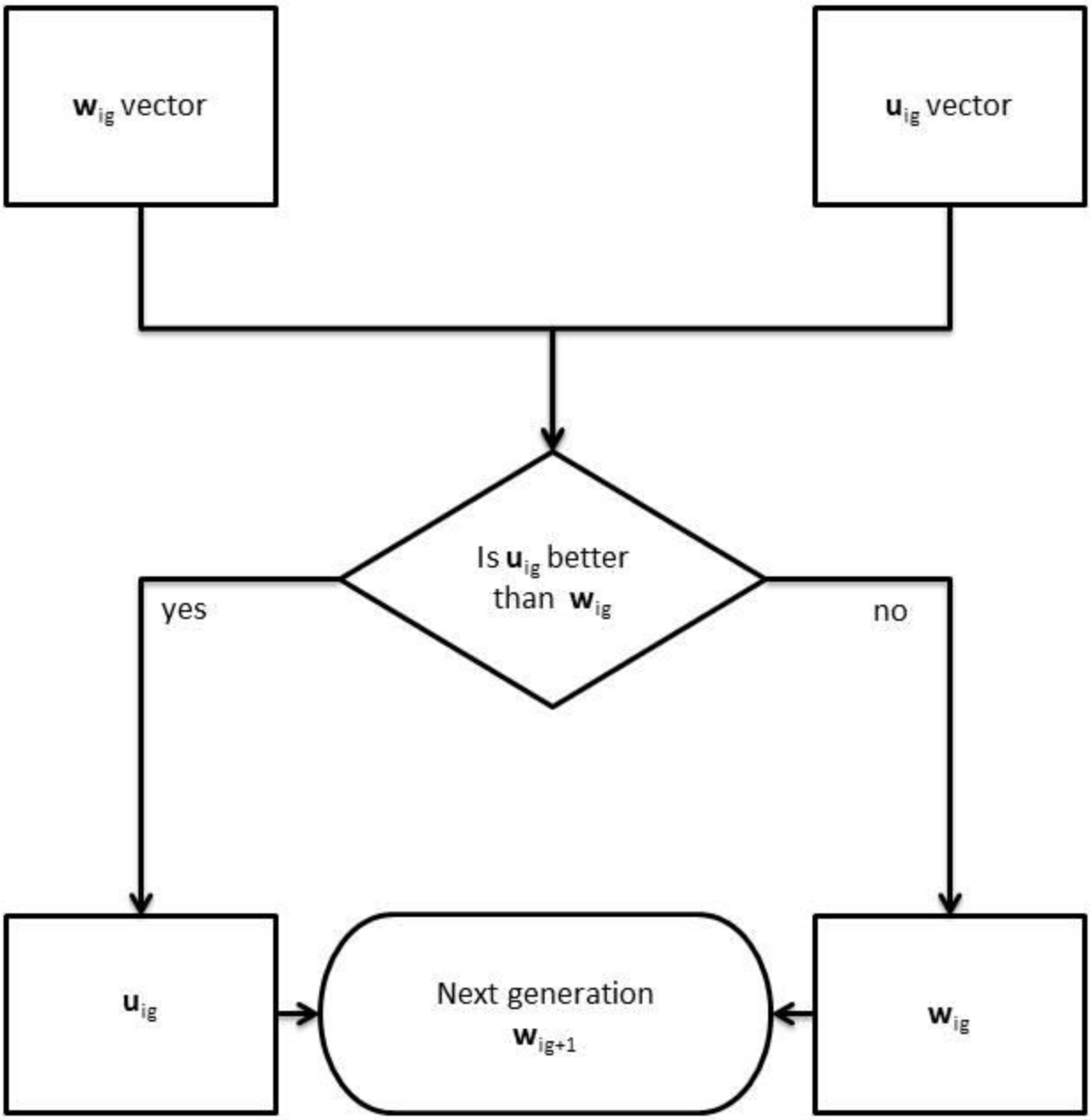
Author contributions: EEF is a coauthor (co-first author), developed the computer code for the kernel mapping and differential evolution, and assisted JJH in the methods development; AB is a coauthor (co-first author) and developed the similarity metrics based on principal component analysis; TAS is a coauthor and provided epidemiologic expertise; SE is a coauthor and assisted in the methods development; JJH is the senior author, conceived of the plan and methods for synthetic data generation and developed computer code.

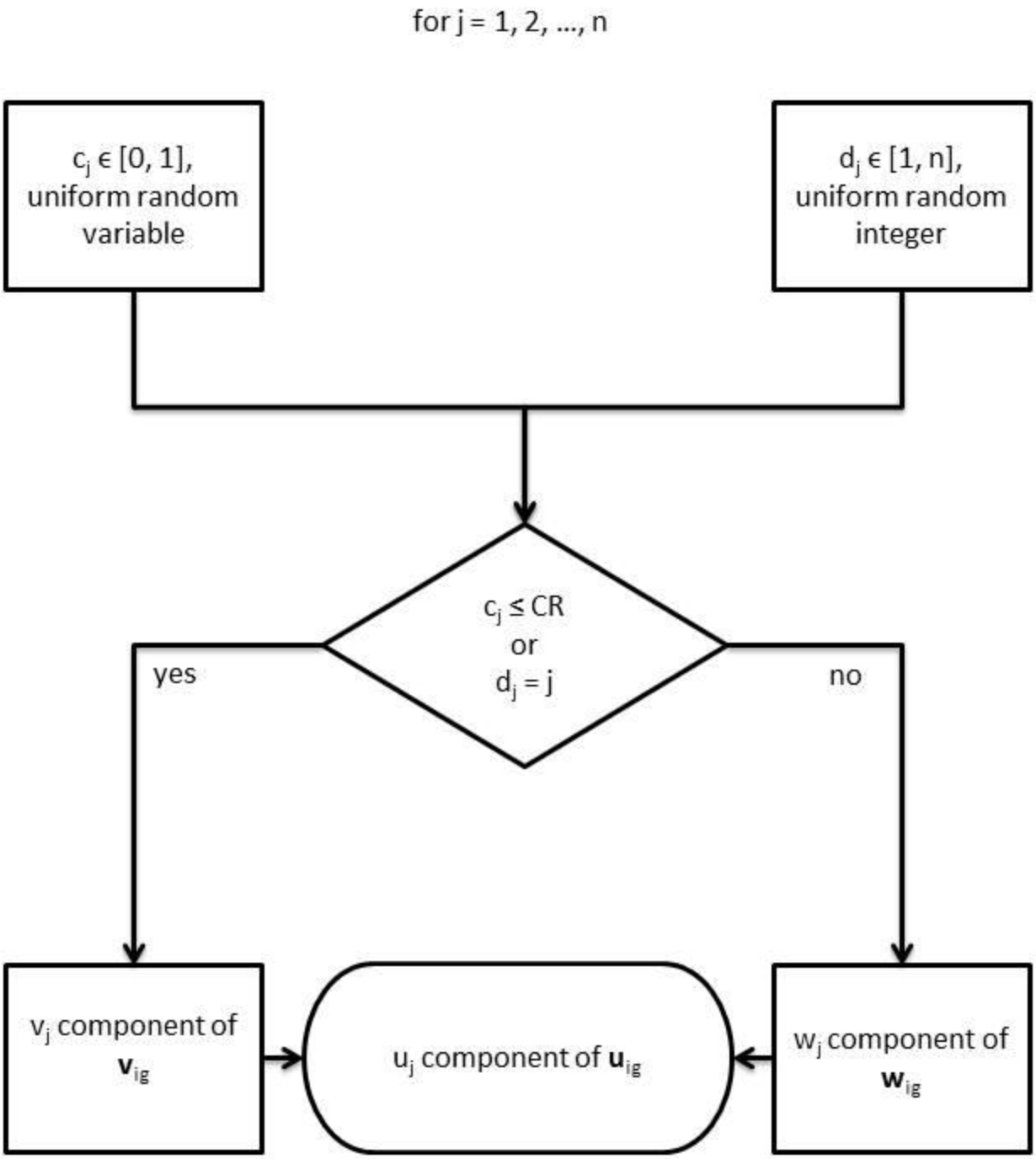
References

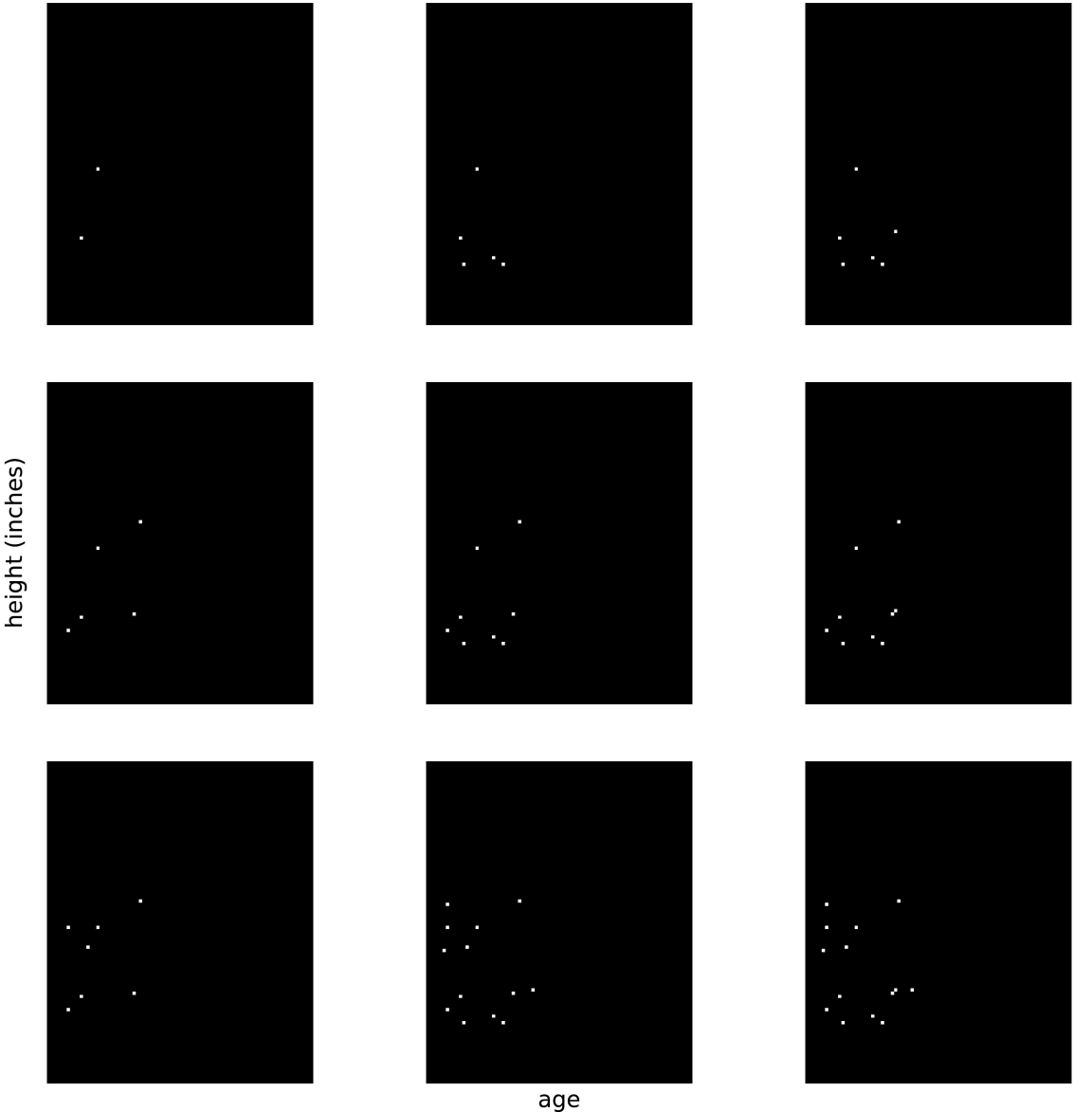
1. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic medicine. 2004;66(3):411-21.
2. Micheel CM, Nass SJ, Omenn GS. Evolution of Translational Omics:: Lessons Learned and the Path Forward: National Academies Press; 2012.
3. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996 Feb 28;15(4):361-87. PubMed PMID: 8668867.
4. Harrell Jr FE. Regression Modeling and Validation Strategies 1997 [updated June, 1997]. Available from: http://biostat.mc.vanderbilt.edu/wiki/pub/Main/ClinStat/model.pdf.
5. Hawkins DM. The problem of overfitting. Journal of chemical information and computer sciences. 2004;44(1):1-12.
6. Efron B, Tibshirani R. Improvements on cross-validation: the 632+ bootstrap method. Journal of the American Statistical Association. 1997;92(438):548-60.
7. Heppenstall A, Harland K, Smith D, Birkin M. Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. Geocomputation 2011 Conference Proceedings, UCL, London2011. p. 1-8.
8. Smith DM, Clarke GP, Harland K. Improving the synthetic data generation process in spatial microsimulation models. Environment and Planning A. 2009;41(5):1251-68.
9. Barthelemy J, Toint PL. Synthetic population generation without a sample. Transportation Science. 2013;47(2):266-79.
10. Müller K, Axhausen KW, Axhausen KW, Axhausen KW. Preparing the Swiss Public-Use Sample for generating a synthetic population of Switzerland: Eidgenössische Technische HochschuleZürich, IVT, Institute for Transport Planning and Systems; 2012.
11. Zhu Y, Ferreira J. Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. Transportation Research Record: Journal of the Transportation Research Board. 2014 (2429):168-77.
12. Harland K, Heppenstall A, Smith D, Birkin M. Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. Journal of Artificial Societies and Social Simulation. 2012;15(1):1-15.
13. Ryan J, Maoh H, Kanaroglou P. Population synthesis: Comparing the major techniques using a small, complete population of firms. Geographical Analysis. 2009;41(2):181-203.
14. Lomax N, Norman P. Estimating population attribute values in a table:“get me started in” iterative proportional fitting. The Professional Geographer. 2016;68(3):451-61.
15. Simpson L, Tranmer M. Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software. The professional Geographer. 2005;57(2):222-34.
16. Ma L, Srinivasan S. Synthetic Population Generation with Multilevel Controls: A Fitness-Based Synthesis Approach and Validations Computer-Aided Civil and Infrastructure Engineering 2015;30:135-50.
17. Williamson P, Birkin M, Rees P. The estimation of population microdata by using data from small area statistics and samples of anonymised records. Environment and Planning A abstract. 1998;30(5):785-816.
18. Voas D, Williamson P. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. Population, Space and Place. 2000;6(5):349-66.
19. Price KV, Storn RM, Lampinen JA. Differential evolution : a practical approach to global optimization. Berlin ; New York: Springer; 2005. xix, 538 p. p.
20. Fowler EE, Lu B, Heine JJ. A comparison of calibration data from full field digital mammography units for breast density measurements. Biomedical engineering online. 2013;12:114. PubMed PMID: 24207013. Pubmed Central PMCID: 3829208.
21. Heine JJ, Cao K, Rollison DE. Calibrated measures for breast density estimation. Acad Radiol. 2011 May;18(5):547-55. PubMed PMID: 21371912. Epub 2011/03/05. eng.
22. Heine JJ, Cao K, Rollison DE, Tiffenberg G, Thomas JA. A Quantitative Description of the Percentage of Breast Density Measurement Using Full-field Digital Mammography. Acad Radiol. 2011 May;18(5):556-64. PubMed PMID: 21474058. Epub 2011/04/09. eng.
23. Heine JJ, Fowler EEE, Flowers CI. A comparison of calibrated and non-calibrated breast density measurements with full field digital mammography Acad Radiol. 2011;18:1430-6.
24. Cacoullos T. Estimation of a multivariate density. Annals of the Institute of Statistical Mathematics. 1966;18(1):179-89.

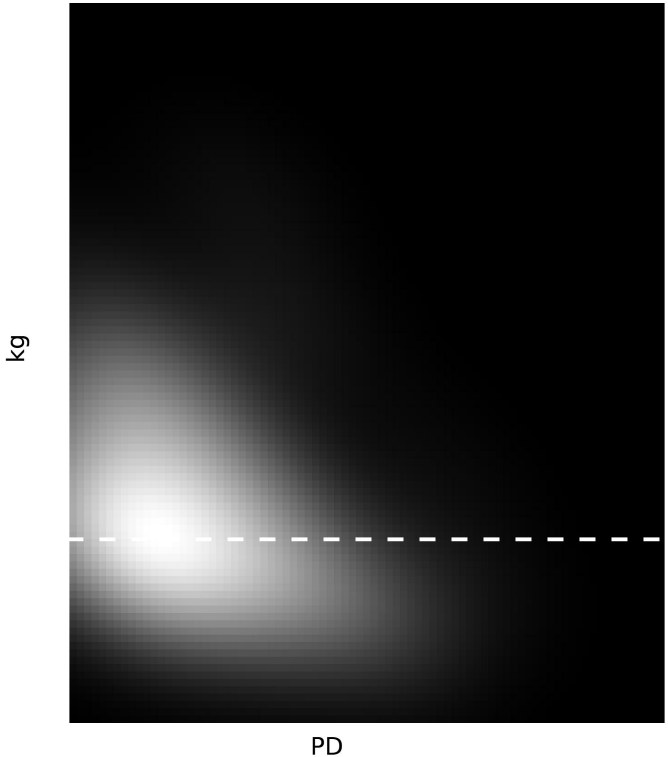
25. Gramacki A. Nonparametric kernel density estimation and its computational aspects: Springer 2018. pages cm p.
26. Gramacki A, Gramacki J. FFT-Based Bandwidth Selector for Multivariate Kernel Density Estimation 2016 12 May, 2016;arXiv:1511.07482v3 [Stat.CO].
27. Duong T, Hazelton M, L. Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation. Scandinavian Journal of Statistics. 2005;32:485–506.
28. Morel P. Gramm: grammar of graphics plotting in Matlab. Journal of Open Source Software. 2018;3(23).
29. Scott DW. Multivariate density estimation : theory, practice, and visualization. New York: Wiley; 1992. xii, 317 p. p.
30. Vernizzi G, Nakai M. A Geometrical Framework for Covariance Matrices of Continuous and Categorical Variables. Sociol Method Res. 2015 Feb;44(1):48-79. PubMed PMID: WOS:000350160200002. English.

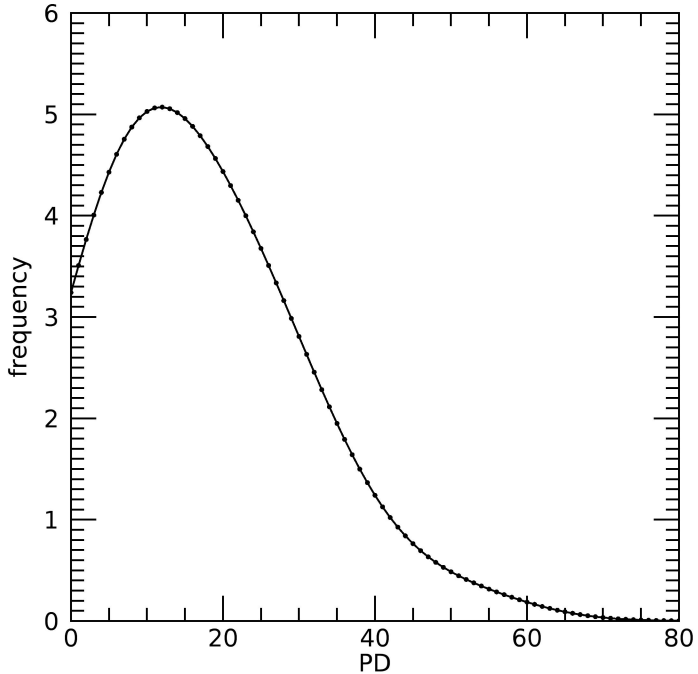
for i = 1, 2, ..., Np

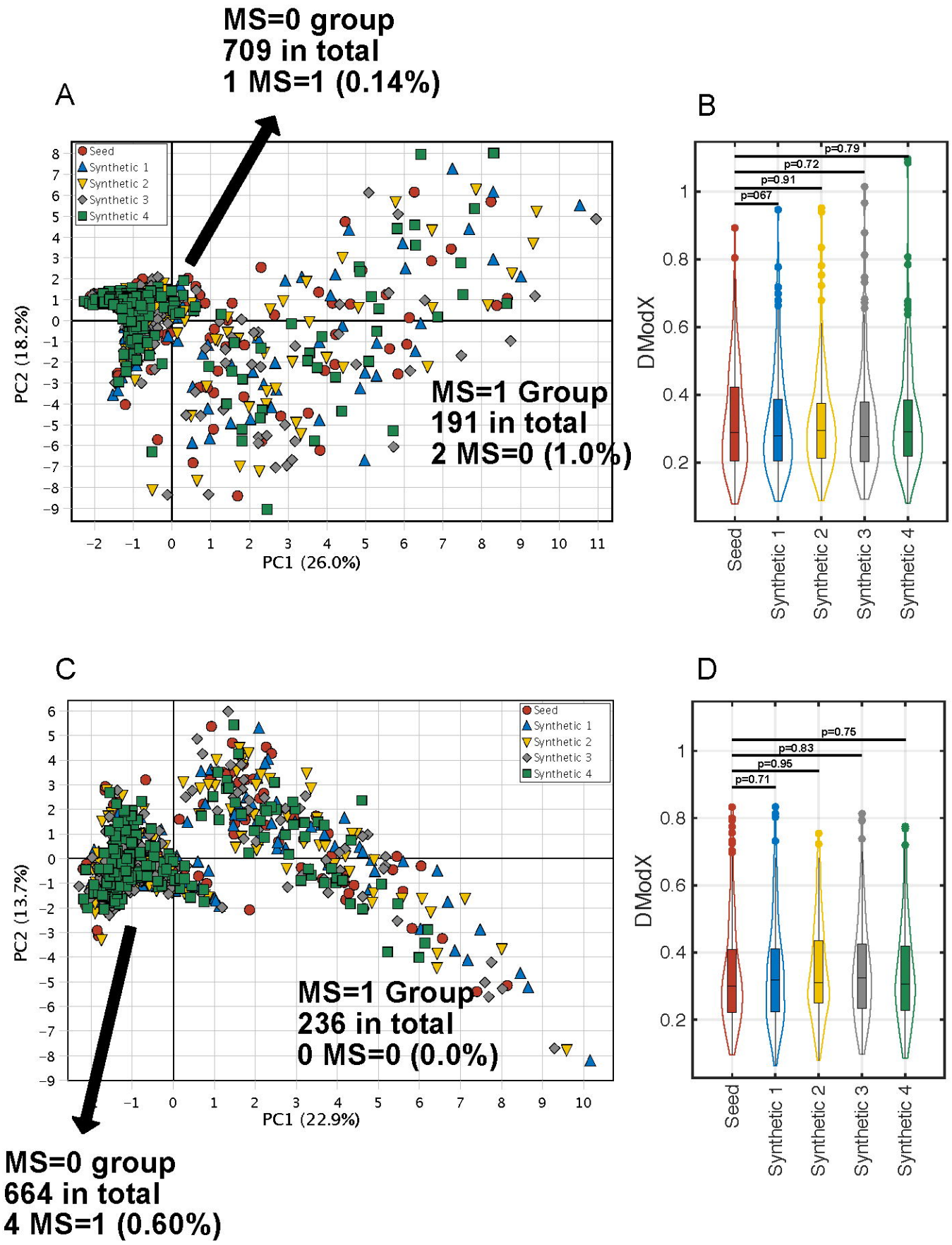

$\mathbf{w}_{ig}$ vector
$\mathbf{u}_{ig}$ vector
Is $\mathbf{u}_{ig}$ better than $\mathbf{w}_{ig}$
yes
no
$\mathbf{u}_{ig}$
Next generation
$\mathbf{w}_{ig+1}$
$\mathbf{w}_{ig}$


for j = 1, 2, ..., n
$c_j \in [0, 1]$, uniform random variable
$d_j \in [1, n]$, uniform random integer
$c_j \leq CR$ or $d_j = j$
yes
no
v_j component of $\mathbf{v}_{ig}$
u_j component of $\mathbf{u}_{ig}$
w_j component of $\mathbf{w}_{ig}$


height (inches)
age


kg
PD


frequency
PD
0
1
2
3
4
5
6
0
20
40
60
80


MS=0 group
709 in total
1 MS=1 (0.14%)
A
Seed
Synthetic 1
Synthetic 2
Synthetic 3
Synthetic 4
PC2 (18.2%)
PC1 (26.0%)
MS=1 Group
191 in total
2 MS=0 (1.0%)
B
p=0.79
p=0.72
p=0.91
p=067
DModX
Seed
Synthetic 1
Synthetic 2
Synthetic 3
Synthetic 4
C
Seed
Synthetic 1
Synthetic 2
Synthetic 3
Synthetic 4
PC2 (13.7%)
PC1 (22.9%)
MS=1 Group
236 in total
0 MS=0 (0.0%)
MS=0 group
664 in total
4 MS=1 (0.60%)
D
p=0.75
p=0.83
p=0.95
p=0.71
DModX
Seed
Synthetic 1
Synthetic 2
Synthetic 3
Synthetic 4