This research presents a data simulation tool designed to provide large-scale meaningful data for classification algorithms (e.g., random forests). Using Python’s pandas, NumPy, SciPy, and random libraries, each simulation generates an unlimited number of unique datasets derived from user-specified global parameters, which include constraints for sample size, the numbers of predictors and the degree of statistical noise, as well as the range of distributional characteristics for the predictors. Sample size is partitioned into two or more categorically dependent outcomes. Dataset characteristics are defined via Python’s pseudo-random number generation. To emulate the natural covariation among predictors in real datasets, the tool transforms initial unadjusted predictor values using randomly specified covariance matrices. This research first details the algorithm. It then demonstrates its potential uses.

### Purpose & Research Questions

**Primary Purpose:** To provide methodologists a tool for exploring ML research questions that demand large-scale data assemblies.

**Secondary Purpose:** To provide an educational tool for those learning about supervised ML algorithms.

**Research Question 1:** Can the algorithmic logic of this approach to simulation properly generate an assemblage of datasets defined by a global parameter space?

**Research Question 2:** To what extent do these simulated data emulate real data from real systems?

---

### User Front-End

**EXECUTE PROGRAM**

1. **GENERATE PARAMETERS**
   - For each dataset, randomly generate:
     - Total sample size
     - Number of possible outcomes
     - Degree of statistical noise

2. **PARTITION OUTCOMES**
   - In each outcome, randomly select:
     - Outcome-specific sample size
     - Random proportion of sample size
     - Outcome-specific noise

3. **SPECIFY CORRELATION MATRIX**
   - For each outcome:
     - Determine size of matrix $M$ where:
       \[ M_{ij} = \rho_{ij} \]

4. **EXECUTE PROGRAM**
   - For each dataset:
     - Determine size of matrix $M$:
       \[ M_{ij} = \rho_{ij} \]

5. **FOR SAMPLE DATASET**
   - Randomly Generated Uncorrelated Predictors
   - Correlated Predictors After Cholesky Decomposition
   - Correlation Coefficients for Transformed Predictors

---

### Future Development

**Forthcoming developments:**
- Graphical User Interface (GUI)
- Questions for consideration:
  - in what other manners may distributional characteristics of predictors vary between outcome groups?
  - What are alternative ways to conceptualize the simulation of noise?

---

### Usage Example

The following example of the usefulness of this data simulation tool is a cross-section of a study currently in progress:

#### Research Questions

1. To what extent can the predictive performance of random forest classifiers be improved by optimizing the ratio of training outcomes?

2. Is it possible to validate the Pareto Principle (i.e., 80/20 rule) that is widely used in classification problems?

3. Does there exist an average optimal t/r ratio parameter for the infinite set of all hypothetical datasets?

#### Rational for large-scale data simulation

Research Questions 2 & 3 cannot be investigated with standard, small-scale approaches to data simulation. They require an immense vector of unique datasets. With this tool, the accuracy and generalizability of the investigation’s findings improve as the number of unique datasets approaches infinity.

---

**Model accuracy as a function of t/r ratio for one dataset:**

![Graph showing model accuracy at different t/r ratios]

Imagine what we could learn about the t/r ratio if we rerun this optimization algorithm with one million more datasets.

---

### References & Author

Matthew C. Myers, M.Ed.

University of Delaware

---

**About the Author**

Matthew Myers is an Associate Professor of Political Science at the University of Delaware. His research primarily focuses on the role of digital technology in politics, with a specific emphasis on social media. He is also interested in the development of data simulation tools for the social sciences.

---

**Scanned by**

University of Delaware