Stat
The ISI's Journal for the Rapid Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.100X/sta.0000

Optimal sample planning for system state analysis with partial data collection

Martin Heller*, Jan Hannig[a], Malcolm R Leadbetter [a]

??Please supply received date??;

We develop optimal and computationally practical procedures to minimize uncertainty concerning the presence of dangerous levels of a contaminant within a building when neither replication nor complete data collection is feasible. More generally, we address inference about the state of a finite system when the state is related to information collected over components of the system when only partial data collection is feasible. When there is no correlation between sample locations a simple random sample or maximum a priori trait presence would provide optimal sampling choices. When complicated probability models describe trait manifestation, the need to collect only partial data precludes a full fitting of complicated models and one must rely heavily on prior information naturally leading to a Bayesian approach. Herein, we introduce a computationally efficient heuristic algorithm to simultaneously find optimal sample locations and decision rule parameterizations then show that it drastically outperforms both random selection and maximum a priori methods. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: uncertainty quantification; statistical planning; decision theory

1. Introduction

The work in this manuscript is motivated by a need to determine whether a building has dangerous levels of a contaminant when a complete census of the building is prohibitively expensive. Unfortunately, when the probability model controlling contamination is complicated, a single observation may require a complete census of the building, and model fitting would require multiple replications of the contamination process applied to the building. As replication would be virtually impossible, we must rely on our best insights into how contamination progresses to perform an accurate analysis on a building. We must first decide where the data should be collected, and then construct a decision rule specific to the data we choose to sample.

More generally, this problem entails making inference about the state of a system, where the state is a summary of traits measured on components of the system when it is impossible to measure the trait over all components. When the observable trait is expressed differently across components of the system, or if the trait has an unusual

[a] **Department of Statistics and Operations Research, University of North Carolina - Chapel Hill**
*__Email: maheller@gmail.com__

correlation pattern, then the choice of evaluation sets could have a drastic influence on the quality of inference. Furthermore, when the distribution of traits differs across evaluation sets, it may be of benefit to adapt decision rules to the selected evaluation set.

This more general problem would include a wide number of scenarios including economic status, pollution levels, population health, individual health, education, production and market research. As an example, an inspector may wish to determine whether the proportion of defects in a product batch exceeds a given threshold. Product quality would have a between and within batch correlation, and maybe even time dependent drift. As another example, a doctor may wish to assess the fat percentage of a patient using caliper measurements from key positions. In either case, the choice of sample and decision rule would highly influence the value of conclusions and the measures of uncertainty.

This problem may be considered one of Bayesian experimental design, yet it differs from the standard Bayesian design problem as defined by Lindley (Chaloner & Verdinelli, 1995). The typical problem consists of selecting predictor variables and, perhaps, loadings thereof in a manner which either increases one's understanding of a system or promotes better decisions after observation of the selected sample. The data collection we envision will be a partial collection of predictors for a single experimental unit without observation of the response variable, making it difficult, if not impossible, to improve the model, parameterization or decision rules. In other words, the posterior distribution of model characteristics will play no part in this analysis. Also, we must rely solely on observation, with no ability to control the values of the predictors in any way. The only control we have available is the selection of which predictors will be observed and the decision rule used to predict the value of the response from the observed predictors. Both the sample selection and decision rule must be determined solely from a priori modeling assumptions.

Optimal sample selection and decision rules have been fully developed under several simple probability models. For instance, one may use a prior distribution for the total number of observations within a finite population which possesses a given characteristic. A sample of size K would then be considered a draw from a hypergeometric distribution with observations from either the affected, or unaffected pool. Based on the number of observations with the affected status, one could then look at the posterior distribution and make a probabilistic claim about whether the number of affected in the population is below a desired threshold (Wright, 1992; Grieve, 1994; Wright, 1997).

An alternative approach assumes that the incidence of a trait within a finite population may be described by independent Bernoulli random variables with a prior placed on the probability of possessing the trait. Again, upon drawing a sample one could look at the posterior distribution for the probability of a trait to make conclusions about how the trait presents within the population (Sego et al., 2007). This latter method has been developed as part of the Virtual Sample Plan software (currently VSP 6.0; Matzke et al. 2010) with the ability to calculate the probability that a population is unaffected based upon the observation of a clear sample and determine the sample size necessary to attain a desired certainty that the population is completely clear of the trait when the sample is found to be clear of the trait.

Both the hypergeometric, and independent Bernoulli models are sufficient for many circumstances including an audit of produced goods when the order of production has little influence. However, the body fat measurement problem and the behavior of a contaminant within a building possess strong correlation patterns and are inadequately described by either of these preceding models. When the variable values at system sites are better described by a more complicated model, then using the model to draw conclusions about a sample would lead to better informed decisions. While it may not be possible to completely understand the exact model underlying the variables of concern, it should be possible to find approximate parametric models along with prior distributions for the parameters of interest suitable for making more informed conclusions. These models are not expected to be perfect, but it is far more arbitrary to use simple models for their mathematical convenience, particularly when they are less capable of modeling system behavior and the simplifications yield poor decisions.

It is impossible to provide a complete encyclopedia of useful models, so we shall focus on problems common to all scenarios where inference must be made on a system when only partial information will be collected without replication. As a proof of concept we demonstrate the contributions of this paper with two examples: In Section 5.2 we use a building mock-up to test our algorithm for the binary sample space with perfect testing. Then in Section 5.3 we demonstrate continuous state space with imperfect testing where the structured modeling characteristics are randomly selected.

The rest of the paper is organized as follows. We present the basic mathematical construction of such problems in Section 2 culminating in a constrained optimization problem. A generic computationally efficient heuristic optimization algorithm is described in Section 3 which would provide a good solution to the objective function of Section 2. As the most basic tool, we propose using Monte Carlo methods to evaluate the objective function while decreasing computational burdens in Section 4, followed by innovations of the optimization algorithm in Section 4.2 to some typical state spaces and sample data. For demonstration, we create a simple branching process model for contamination within a building to contrast the efficacy of the sampling approaches in Section 5. We conclude the paper with some comments and suggestions in Section 6.

2. Mathematical presentation

The methods described herein require that a system can be decomposed into a collection $\mathfrak{T}$ of $N < \infty$ components. Any two components should be distinct, and each be sufficiently small so that the appropriate data collection technique would summarize the observable variables over the component. The components do not need to represent identical units with independent presentation of the observable variables. It is not even necessary for each random variable X_t to be on the same space. The observables may be continuous, discrete or categorical, and they may measure very different phenomena. The components could be individual members of a population or non-overlapping geometric regions. A graph of random variables X, indexed by the collection $\mathfrak{T}$, will be used to summarize the true trait values at the components. When interested in the observed values over an evaluation set $T = \{t_1, t_2, \ldots, t_n \in \mathfrak{T}\}$, we shall use the shorthand X_T.

Given the value X_t, we must rely on the observed values $Y_t \sim \phi(X_t, \epsilon_t, t)$, where ϵ_t are independent random variables reflecting measurement error. If Y_t is a perfect rendering of the state, then $Y_t \equiv X_t$. Since the measurement error is based on the measurement technique, it is assumed that the nature of the error is well known, or can be estimated through a controlled calibration.

Our goal is to make an inference about the state S of a system, and we suppose that S is related to the observed values Y_T through a relationship $S = g(Y_T, \epsilon_{Y_T})$.[†] In most cases, it will be impractical to observe all of Y, so we shall consider observations over an evaluation set $T \in \mathfrak{T}$, where $|T| = K$, i.e. the sample size is limited to K collected data points. Inference about S will be based on the observed random variables Y_T, so we wish to select the most informative subset T and given Y_T, make the best inference about S. The inference will be through a decision $d(Y_T, \beta)$, where the parameter β reflects a selected strategy and may depend on the selected T. The quality of the decision will be judged by the estimated performance of a loss function $L(S, d)$, which gives a numerical value for the severity of predicting d when the true state is S. The optimization problem may be succinctly written as

$$\{T^*, \beta_{T^*}\} = \operatorname*{arg\,min}_{T\in\mathfrak{T},\beta\in\Re^k:|T|=K} r(T, \beta), \tag{1}$$

where $r(T, \beta)$ is the Bayes risk associated with the decision $d(Y_T, \beta)$, namely,

$$r(T, \beta) = \int \int \int L(S, d(Y_T, \beta))\, dP_{Y,S|X}\, dP_{X|\theta}\, dP_\theta,$$

where we assume the prior P_θ over the set of potential probability models generating the random characteristics X over the graph nodes.

3. Algorithm for finding an optimal set of testing sites and decision rule

Finding an exact solution to Equation 1 would require excessive computational time. First, simply evaluating $r(T, \beta)$ analytically is expensive and particular to the selected probability models. Second, an exhaustive search would require N^K evaluations (including duplicate elements) of $r(T, \beta)$.

Without performing any calculations, there are two naive solutions which should produce reasonably good results. The first is to select entries by random selection. The second is to select the K members of $\mathfrak{T}$ which each individually attain one of the K smallest losses $\min_{\beta \in \Re^k} r(t, \beta)$. The first method will be succinctly referred to as RAND and the latter as MN. Under the simple hypergeometric or independent Bernoulli distributions either of these methods would perform equally well and as well as any other sample selection mechanism. If a suitable probability model for the spread of contamination can not be envisioned with confidence then either RAND or MN should be considered as viable strategies for data collection. However, when a realistic probability model is available, we will show that our algorithmic approach will consistently outperform either of these strategies.

We make use of a necessary condition which an optimal solution must satisfy to guide the construction of heuristic optimization algorithms. Suppose that T^* minimizes the risk in Equation 1 and satisfies $|T^*| = K$. If $L < K$, we may select subsets $S^* \subset T^*$ and $S' \subset \mathfrak{T}$ each of size L and create a new evaluation set of size K by setting $\tilde{T} = (T^* - S^*) \cup S'$. Since T^* is an optimal evaluation set we would have $\min_{\beta \in \Re^k} r(T^*, \beta) \leq \min_{\beta \in \Re^k} r(\tilde{T}, \beta)$. This necessary condition for a solution suggests an iterative algorithm wherein individual elements of the current evaluation set are replaced with other elements of $\mathfrak{T}$ which decrease the loss. When no improvement is possible through individual replacement, or a termination condition is satisfied, the algorithm ends. The algorithm is as follows:

1. Select a set $T_0 = (t_1, t_2, \ldots, t_K)$ of K positions from $\mathfrak{T}$ as an initial evaluation set.
2. Set i=0.
3. For each of the $j = 1, \ldots, K$ sample positions in the current evaluation set do the following:
   (a) Remove the j-th element from T_i, to create the test set $\tilde{T}_{i,j}$.
   (b) Replace the j-th element of T_i and β_i with the solution of the following:

$$\{t_j^*, \beta\} = \underset{\tau \in \mathfrak{T}, \beta \in \Re^k}{\arg\min}\, r(\tilde{T}_{i,j} \cup \tau, \beta).$$

4. Let T_{i+1} be the final evaluation set of the previous loop and set $i = i + 1$. Return to Step 3 unless a termination condition is satisfied.

[†] The relationship between S and X may be deterministic or random, i.e. $S = h(X)$ or $S = h(X, \epsilon_X)$. This will not make any difference for our set-up since we are relying on the relationship $S = g(Y_T, \epsilon_{Y_T})$ where $T \in \mathfrak{T}$.

This generic algorithm serves as the most basic form for all subsequent optimization algorithms, with adaptations of Step 3b for efficiency and conciseness based on the hypothesized scenario. As the algorithm is based on a necessary condition, any convergent solution will be a local minima in the sense that it can not be improved upon through replacement of individual components. While not part of the present work, evaluation set size calculations could also be performed based on these techniques. First one would perform the optimization for samples of size $K = 1, 2, \ldots$, then find the minimum value of K which achieves the desired loss.

The computational complexity for this algorithm would be equal to $O(RKG)$, where R is the number of times the loop in step 3 needs to be performed, K is the number of replacements which must be tested, and G is the computational complexity of step 3b. Depending on the scenario, the complexity of G will reflect the algorithm used to compute the risk.

While any starting set $\{T_0, \beta_0\}$ will be valid, we will use the following to get a good starting point:

1. Set $\dot{T}_0 = \emptyset$.
2. For $i = 1, \ldots, K$
    (a) Let $\dot{T}_{i+1} = \dot{T}_i \cup t_i^*$, where $\{t_i^*, \dot{\beta}_{i+1}\} = \underset{\tau \in \mathfrak{T}, \beta}{\arg\min}\, r(\dot{T}_i \cup \tau, \beta)$.
    (b) Set $i = i + 1$.
3. Set $\{T_0, \beta_0\} = \{\dot{T}_K, \dot{\beta}_K\}$.

4. Innovations to building contamination

For the contamination problem, we wish to determine whether a building has dangerous levels of a contaminant. The state space for S will take only two values, namely, 1 when $X_t > \beta_0$ for at least one $t \in \mathfrak{T}$, and 0 otherwise, i.e., $S = I(\max X_i > \beta_0)$. A natural loss function would be

$$L(a, b) = I(a = 1)I(b = 0) + \alpha I(a = 0)I(b = 1), \tag{2}$$

where the coefficients differ because there may be greater loss incurred by claiming a building is clear when it is not. Reflecting the state space definition, we will consider decisions of the form $d(Y_T, \beta) = I(\max_{t \in T} Y_t > \beta)$ to predict the state S. While this choice of decision rules may not be optimal for all of the scenarios presented, it will be sufficient to demonstrate how the algorithm works and why there is a benefit to using it. When one selects a decision rule, it is important to consider both the optimality of the rule class along with the computational complexity required to optimize the rule. For our selected decision rule, $E(I(S = 1)I(d(Y_T, \beta) = 0))$ and $E(I(S = 0)I(d(Y_T, \beta) = 1))$ have monotone derivatives with respect to β making an optimal choice easy to find with a search over a suitable bounded interval.

For contamination there are two scenarios which should cover the majority of models based on whether the observed data Y_t are binary or continuous. We present the simplifications for these two scenarios in Section 4.2. As presented, the algorithms will be suitable either when Y is a perfect rendering of X, or when there is measurement error. When Y is binary, we set $\beta = .5$ removing optimization of the decision rule for the algorithm.

4.1. Number of simulated samples needed for accurate analysis

It is necessary to calculate the expectations $r(T, \beta)$. While it may be possible to find an analytic solution, it is likely to be a computationally expensive process and particular to a specific model. For the model presented in Section 5.1 even the calculation of the simple conditional probability $P(Y_{t_1} > \beta | Y_{t_0})$ involves N nested sums of binomial coefficients and polynomials, because the correlation must be tabulated through all possible pathways between nodes t_0 and t_1. These "simple" calculations have computational complexity $O(N!)$ for each formal estimate. Using a non-polynomial time calculation for $r(T, \beta)$ would thwart the computational efficiency gained by our proposed algorithm. Furthermore, the need to construct an analytic solution could require tremendous mathematical and computational ingenuity making a solution far more labor intensive for the practitioner.

Rather than using analytic methods, we suggest relying on Monte Carlo simulation to estimate the values as $\hat{r}(T, \beta)$. Many probability models make simulation trivial to code and the simulations are computationally efficient. When N is fixed, there are no more than N^K possible choices for an evaluation set, T, of size K, and this is a finite number. Since the maximum number of evaluation sets is finite, $L(S, d(Y_T, \beta))$ has a finite shattering number, and integration is continuous; $r(T, \beta)$ defines a Glivenko-Cantelli class meaning that there will be uniform convergence with respect to the sample size, i.e.

$$\max_{T,\beta} \|\hat{r}(T, \beta) - r(T, \beta)\| = o_p\left(\frac{1}{\sqrt{M}}\right),$$

where M is the number of simulated samples. Hence, one may achieve arbitrarily high precision (in probability) for each of the needed estimates $\hat{r}$ simply by increasing the number of simulated sample.

The number of simulated graphs M needed for accurate sample selection will grow as both the number of positions N and the size of the evaluation sets K increase. For our subsequent algorithm, we will need to perform at most $H \leq Q\binom{N^K}{2} = O(N^{2K})$ comparisons of $\hat{r}(T_a, \beta_a)$ vs $\hat{r}(T_b, \beta_b)$, where Q is a constant related to the precision needed for β independent of both N and K. Suppose we want to distinguish a difference of $\epsilon > 0$ with the probability of a single mistake over all H comparisons no more than α. This precision would be assured (asymptotically) if we simulate at least

$$\mathcal{M} = 4\left(\frac{z^*_{\alpha/H}}{\epsilon}\right)^2 \sup_{T \in \mathfrak{T}, \beta : |T| = K} var\{\hat{r}(T, \beta)\}$$

samples, where z^*_p is the $p-$th quantile of the standard normal distribution function and α/H is the Bonferroni correction for multiple comparisons. For sufficiently small values of p, $z^*_p \propto \sqrt{\log(1/p)}$ (Dominici, 2003). Replacing p with α/H we find that $\mathcal{M} \propto \log(H/\alpha) \leq O(K \log N)$.

Subsequent applications assume that one simulates M replications of the random graph Y. These replications are stored in an $M \times N$ matrix $\mathbb{Y}$ where each row stores a single simulated graph. The storage of the matrix $\mathbb{Y}$ (and other matrices of the same size derived from this) will be the major burden on the storage needed for computation making the storage requirements of order $O(NM) \leq O(NK \log(N))$. It is useful to note that given the simulated values of $\mathbb{Y}$, the algorithm presented previously makes no use of the probability model used to generate the data. If explicit formulas for $r(T, \beta)$ were used, the model could play a more prominent roll in algorithm development.

It should also be noted that since only the matrix of generated data $\mathbb{Y}$, and the actual state S are needed, one may substitute real data and completely forego the need to construct probability models. This could provide insight when one is able to conduct extensive studies on a number of systems to try to find the most appropriate test sites for minimum testing costs in later studies of similar systems.

4.2. Specific optimization algorithms

The innovation of the algorithm using simulated data is quite similar to the Federov Exchange Algorithm (FEA) (Fedorov, 1972). The FEA exchanges the rows (predictor values) of a design matrix Y to minimize the variation in estimators, for instance, the trace of the information matrix $Y'Y$. The performance is optimized by executing only the necessary calculations to see how replacement of one row with another would affect the objective function. In contrast to the FEA, our optimization is with respect to column selection (variable/location selection). Herein, we include the modifications of step 3 of the algorithm presented in Section 3 to improve performance. The idea for the binary decision is to remove a test location, t_i, then find the improvement attained by adding a new test cite t_j using the minimal amount of calculations necessary, i.e. $\hat{r}(T - t_i + t_j, .5) - \hat{r}(T - t_i, .5)$. Starting with $j = 1$, the algorithm may proceed as follows. (The computational complexity of each step is presented in square brackets.)

(a) Remove the j-th column from the evaluation data matrix, i.e. set $\tilde{\mathbb{Y}} = \mathbb{Y}_{T_i - t_j}$, where $T_i - t_j$ refers to column indices and t_j is the j-th element of T_i. [$O(KM)$]
(b) Let $\mathbb{Y}_0$ and $\mathbb{Y}_1$ be the rows of $\mathbb{Y}$ where no contamination is detected in the rows of $\tilde{\mathbb{Y}}$ and either $S = 0$ or $S = 1$, respectively. (If contamination is detected in one of the $\tilde{\mathbb{Y}}$ columns, then the observed values in the added column will not change the decision for that replication.) [$O(NM)$)]
(c) Sum each column of $\mathbb{Y}_i$ to create a vector C_i for $i \in \{0, 1\}$. [$O(NM)$]
(d) Set $B = \alpha * C_0 - C_1$. (This is a row vector with entry j equal to $M[\hat{r}(T - t_i + t_j, .5) - \hat{r}(T - t_i, .5)]$.) [$O(NM)$]
(e) Find the index j^* of B with the smallest value. [$O(N)$]
(f) Replace the j-th entry of T_i with j^*. [$O(K)$]
(g) If $j = K$ quit, otherwise set j=j+1 and return to (a). [$O(1)$]

We see that the inner loop is composed of sequential $O(NM)$ terms, making the full algorithm $O(RKNM) \leq O(RNK^2 \log N)$. While K and N will be fixed values, the behavior of R will depend on the situation and there is no clear relationship between K,N and R.

The algorithm for the continuous observable scenarios may proceed exactly as in the general algorithm of Section 3 with r replaced by $\hat{r}$. The only thing which must be established is the optimization of β once the evaluation set T is selected. For our decision rule, $d(Y_T, \beta) = I(\max_{t \in T} X_t > \beta)$, we could select the lower bound $\beta_0 = \overline{\max(Y)}$ for $S = 0$, and upper bound $\beta_1 = \overline{\max(Y)}$ for $S = 1$. Optimization of β will be performed by a golden ratio search over the range $[\beta_0, \beta_1]$. Depending on the desired precision, the search will require $Q = O(1)$ operations, where Q is completely independent of N, M, and K. The evaluation of each $\hat{r}(T, \beta)$ would be of order $O(KM)$ and it would need to be performed for each of the N potential testing sites, making the combined complexity $O(RK^2NM) \leq O(RNK^3 \log(N))$.

5. Simulations

We now present simulations of both a binary state space with perfect testing (BPT), and a continuous state space with imperfect testing (CIT) where the random graphs will be simulated using a generative probability model presented in Section 5.1. For each simulation we will create multiple training sets of fixed sizes, M, and find evaluation sets using either ALG or MN on the training sets. The selected evaluation sets will be tested using a much larger test set generated independent of the training sets. We will test the RAND method by randomly selecting evaluation sets of size K and using the large test set to estimate the risk. The loss function for both scenarios will be of the form in Equation 2 and use $\alpha = .2$.

5.1. *Probability model used for simulation*

Our model supposes that contamination starts at a single seed location $\dot{t}$ and then spreads to other graph nodes by a branching process constrained to the graph $\mathfrak{T}$. The initial position, $\dot{t}$ will be selected using a random draw from a multinomial distribution with weights corresponding to the relative probability of contamination starting at a given location, including the category $\emptyset$ meaning that no seed is selected hence, no contamination is present. Contamination will spread based on an $N \times N$ matrix of probabilities, Π, where the i, j-th element denotes the probability of contamination spreading directly from position i to position j when position i is contaminated and j is currently clear. If the values of Π are high, then it is likely that all positions within the graph will become contaminated making simulation unexciting. If the values of Π are sufficiently small, then the branching process will terminate prior to complete saturation with high probability. If the values of Π are set to be too small, then contamination will not spread and the resulting simulations will yield results very similar to the seed contamination distribution.

Begin by setting the initial contamination set to $E_0 = \{\dot{t}\}$, and the susceptible set to $S_0 = \{\mathfrak{T} - \dot{t}\}$. Also set $X_{\dot{t}} = 1$ for the binary state space and $X_{\dot{t}} \sim Exp(\lambda)$ for the continuous state space. From the seed location, the model will propagate for the binary scenario through the following algorithm:

```
**While** $E_i \neq \emptyset$ and $S_i \neq \emptyset$
    **Set** $E_{i+1} = \emptyset$ and $S_{i+1} = S_i$.
    **For** $e_j \in E_i$
        **for** $s_j \in S_i$
            **Draw** $u \sim U(0, 1)$.
            **if** $u \geq \Pi_{i,j}$
                **Remove** $s_j$ from $S_{i+1}$
                **Add** $s_j$ to $E_{i+1}$
                $*$**Set** $X_{s_j} = 1$
```

For the continuous case, we will replace the starred step with $X_{s_j} \sim Exp(\lambda)$. When imperfect testing is expected, the observed values Y_{s_j}, will be generated from the matrix X_{s_j} using the appropriate random error.

5.2. *Binary sample space with perfect testing (BPT)*

We will use a building mock-up to test our algorithm for the binary sample space with perfect testing. The building will have 4 identical floors with 19 rooms each as depicted in Figures 1(a-c). Of these 76 locations, we intend to find 10 sites which minimize the risk $r(T, \beta)$. The numbering scheme (shown for the first floor) was selected since it will impose structures on portions of the propagation effects matrix as depicted in Figure 1d.

The prior assumes a 90% chance that contamination will be present. Figure 1(a) displays the between and within hallway effects. The arrows connecting adjacent rooms on the same side of the hall will alter the A components of the propagation matrix as

$$A = \begin{bmatrix} 1 & .35 & 0 & 0 \\ .35 & 1 & .35 & 0 \\ 0 & .35 & 1 & .35 \\ 0 & 0 & .35 & 1 \end{bmatrix}. \tag{3}$$

The opposite hallway effects are contained in the B submatrices and expressed as $B = .2I_4$, where I_Q is an identity matrix of dimension Q. Likewise, the between wall effects would be contained in the C component and be $C = .005I_4$. Since the bathrooms are single sex, we can assume that contamination from or to the bathroom corresponds to the sex of an office occupant. Given the gender of office workers in Figure 1(b), submatrix D will be set to

$$D = .15 \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \cdots \end{bmatrix}.$$

Assuming that each individual interacts with the office in much the same way, would mean that E is a row vector with identical values, which we set to $.2$. Transmission between floors will be through 30 randomly selected pairs on separate floors with the probability of transmission as a random draw from the beta distribution with parameters (1,5).

The results of this scenario are presented in Figure 2. It is clear that the algorithm provides much better results than either the MN, or RAND solutions, while attaining much less variance. In fact, within each sample size, the worst solution attained through the algorithm was better than the best solution attained by MN, and RAND. Over all simulations, the maximum number of iterations (R) needed were 4, and the mean number was around 1.9 for all training set sample sizes with no discernable trend.

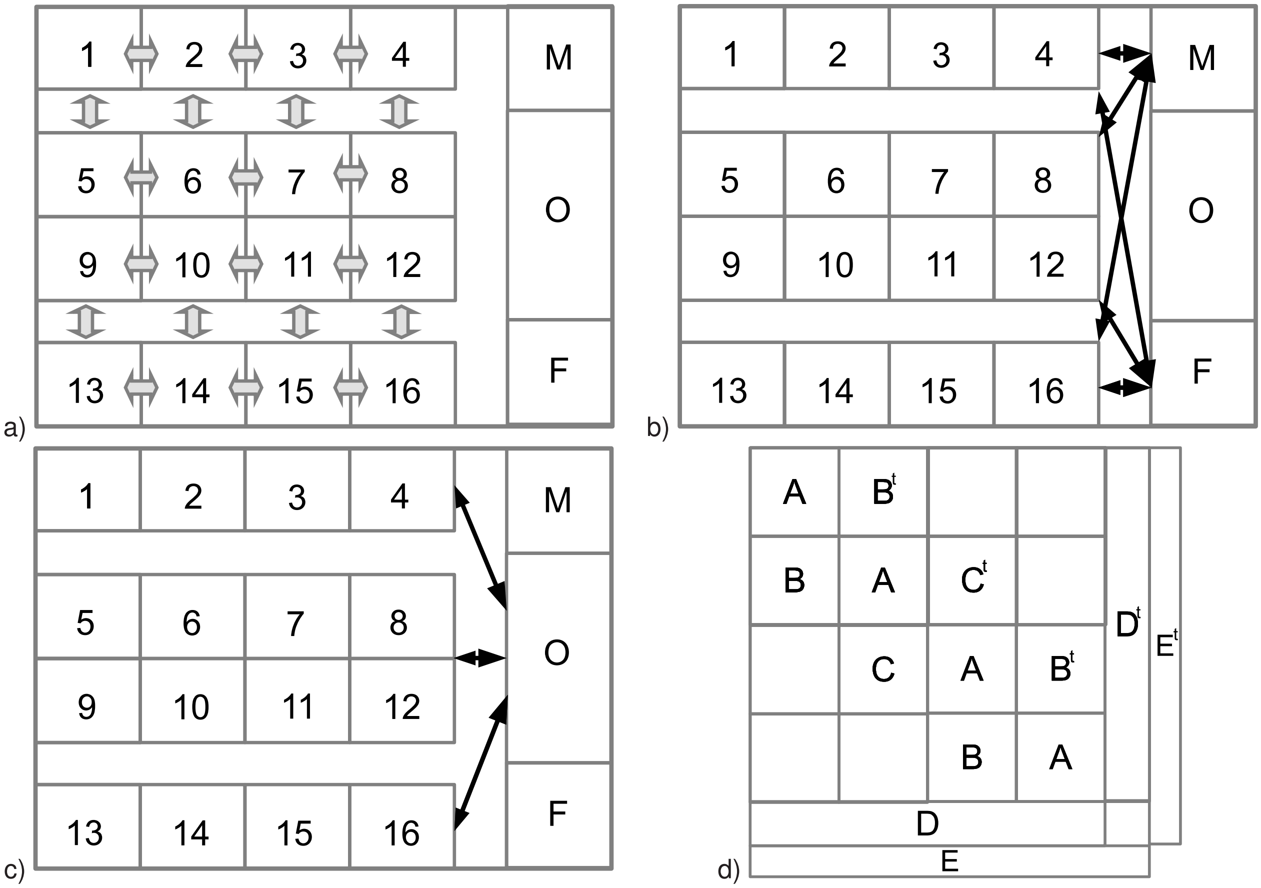

Figure 1. Several effects (a) within and between hall (b) bathroom (c) main office (d) submatrix structure

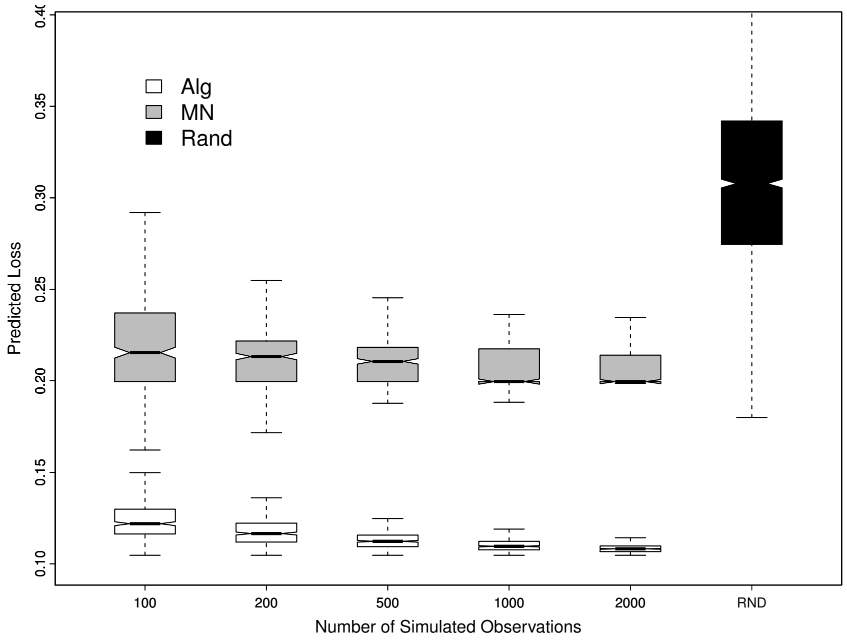

Figure 2. Estimated loss for binary observable with perfect testing method

5.3. Continuous state space, imperfect testing (CIT)

For this scenario we simulated a random matrix of $N = 64$ sites with the goal to find the $K = 18$ sites which would maximize the objective function. The propagation matrix Π was created by first dividing the 64 sites into 5 groups by randomly placing each site into one of the 5 groups with equal probability. The symmetric propagation matrix was filled using

$$\Pi_{i,j} = \begin{cases} I(u_{i,j} > .2)z_{i,j} & i,j \text{ in same group} \\ I(u_{i,j} > .5)\tilde{z}_{i,j} & i,j \text{ in different groups,} \end{cases}$$

where $u_{i,j} \overset{iid}{\sim} U(0,1)$, $z_{i,j} \overset{iid}{\sim} \beta(4,2)$, and $\tilde{z}_{i,j} \overset{iid}{\sim} \beta(1,5)$. After the propagation matrix was created, it was used for every simulated observation in both the training sets and the test set. The seed was selected by randomly selecting 5 of the 64 sites then selecting one of them with equal probability for each simulated graph. When contamination was transferred to a site, the concentration was randomly selected using $Exp(.5)$. We then concluded that a site was contaminated if it exceeded the .92-quantile of the $Exp(.5)$ distribution function. The measurement error was set to $y_i \overset{iid}{\sim} N(x_i, x_i/4)$.

We created a test set of 150000 simulated random graphs to test the achieved clearance probabilities. The training sets were created as separate sets of size $M = (500, 1000, 2000, 5000, 10000)$ random graphs. For each training set size, we performed the simulation 1000 times to produce the results of Figure 3. Surprisingly, the RAND and MN methods performed nearly identically for this model with about twice the loss of the ALG solution, and possessed considerably greater variance. The number of iterations, R, needed for ALG did not exceed 7, and only needed 7 iterations one time throughout all of the simulations. The mean number of iterations needed was 2.64, and counterintuitively decreased from 2.83 down to 2.56 as the number of random graph replicates in the training sample increased. To verify this trend, we ran the experiment multiple times and found the decrease to be typical.

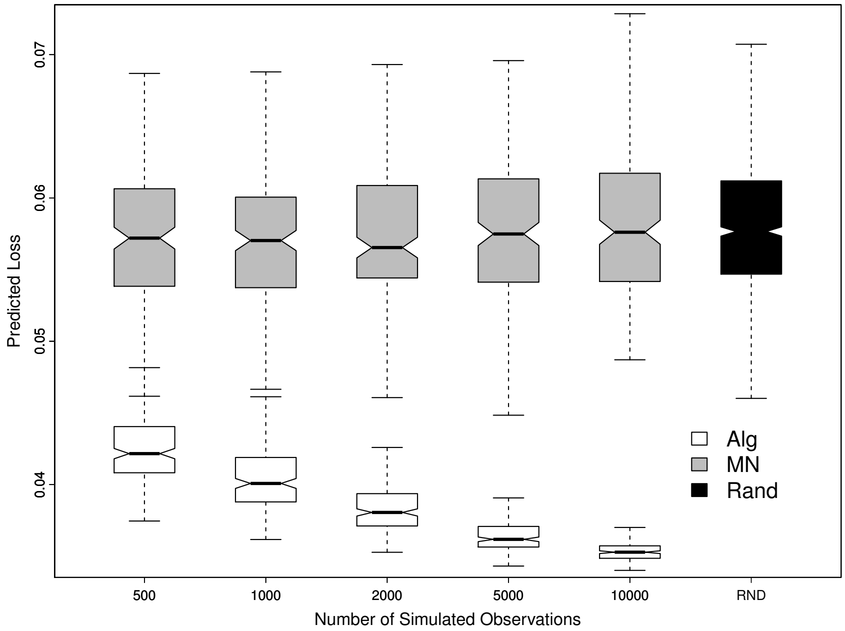

Figure 3. Estimated loss for continuous observable with imperfect testing method

6. Conclusion

We have shown that our proposed algorithmic approach to finding optimal samples and decisions would considerably decrease the expected loss when a suitable probability model is available. The algorithm seems to be quite efficient with maximum computational order of $O(RK^2N\log(N))$, and $O(RK^3N\log(N))$ for our two implementations. While R can not be bounded by a polynomial term of N and K, it is not expected to be very large. If it is a problem, a suitable stopping condition should be added to maintain efficiency.

The gain realized from sample selection is based on the variation of

$$f(T) = \inf_{\beta \in \Re^k} r(T, \beta)$$

over sets T of a fixed size. If the prior information is too weak, the variation of $f(T)$ may be small making little improvement possible by sample selection. Still, one should use the best available prior distribution to estimate the risk $r(T, \beta)$.

The core algorithm we presented could be useful in a wide variety of model structures. As an example, suppose the state is related to the component values through a linear regression $S_i = \beta' X_i + \epsilon_i$ and we use a squared error loss. Then the algorithm could be used to find the subset of size K which achieves the minimum risk and the optimization for the decision rule would then be a standard regression analysis.

The cost of sampling of the set T could be generalized to $C(T)$ and the constraint could be $C(T) \leq Q$. In this case, one would need to test replacement of a single observation with any subset which does not cause $C(T)$ to exceed Q. While this approach might not have nice computational bounds, it should reduce the computational burden of execution drastically compared to the exhaustive approach.

The main hurdle for this method is the need to construct models for the system of interest. In our simulations, the prior distribution used only a single generative model. In reality, a more fully Bayesian setup would be expected with priors placed on the model parameters or even the model structure thus weakening the prior information. Even with this extension, it may be unlikely that the true model is in the support of the prior. Fortunately it is possible to gain insights on phenomena even when a crude model is used. If one is uncomfortable with their modeling choices, they may use the algorithm to select the "best" sites then proceed as if the sample were drawn from a hypergeometric, or a binomial distribution. Alternatively, one could use a weighted average of the loss attained from the simpler and more complicated probability models to draw conclusions.

In short, this algorithm is quite powerful when one is able to confidently build models to describe the behavior of system components. Even when the modeling assumptions are not precisely satisfied, the efforts spent providing more specific models should provide a better understanding of decision uncertainty than would be attained by using a simpler less applicable model.

References

Chaloner, K & Verdinelli, I (1995), 'Bayesian experimental design: A review,' *Statistical Science*, **10**(3), pp. 273–304.

Dominici, DE (2003), 'The inverse of the cumulative standard normal probability function,' *Integral Transforms and Special Functions*, **14**, pp. 281–291.

Fedorov, V (1972), *Theory of Optimal Experiments*, New York, NY: Academic Press.

Grieve, A (1994), 'A further note on sampling to locate rare defectives with strong prior evidence,' *Biometrika*, pp. 787–789.

Matzke, B, Wilson, J, Nuffer, L, Dowson, S, Hathaway, J, Hassig, N, Sego, L, Murray, C, Pulsipher, B, Roberts, B & McKenna, S (2010), *Visual Sample Plan Version 6.0 User's Guide*, Pacific Northwest National Laboratory, pnnl-19915 edn., richland, Washington.

Sego, L, Anderson, K, Matzke, B, Sieber, W, Shulman, S, Bennett, J, Gillen, M, Wilson, J & Pulsipher, B (2007), 'An environmental sampling model for combining judgment and randomly placed samples,' Tech. Rep. PNNL-16636, Pacific Northwest National Laboratory, richland, WA.

Wright, T (1992), 'A note on sampling to locate rare defectives with strong prior evidence,' *Biometrika*, **79**, pp. 685–691.

Wright, T (1997), 'A simple algorithm for tighter exact upper confidence bounds with rare attributes in finite universes,' *Statistics & Probability Letters*, **36**(1), pp. 59 – 67.