UNIVERSITY OF
Southampton
Southampton Statistical
Sciences Research Institute

Weighted space filling designs for dependent variables with application to deterministic computer codes

M. Feddag[†⋆], D.C. Woods[†] and V.E. Bowman[‡]

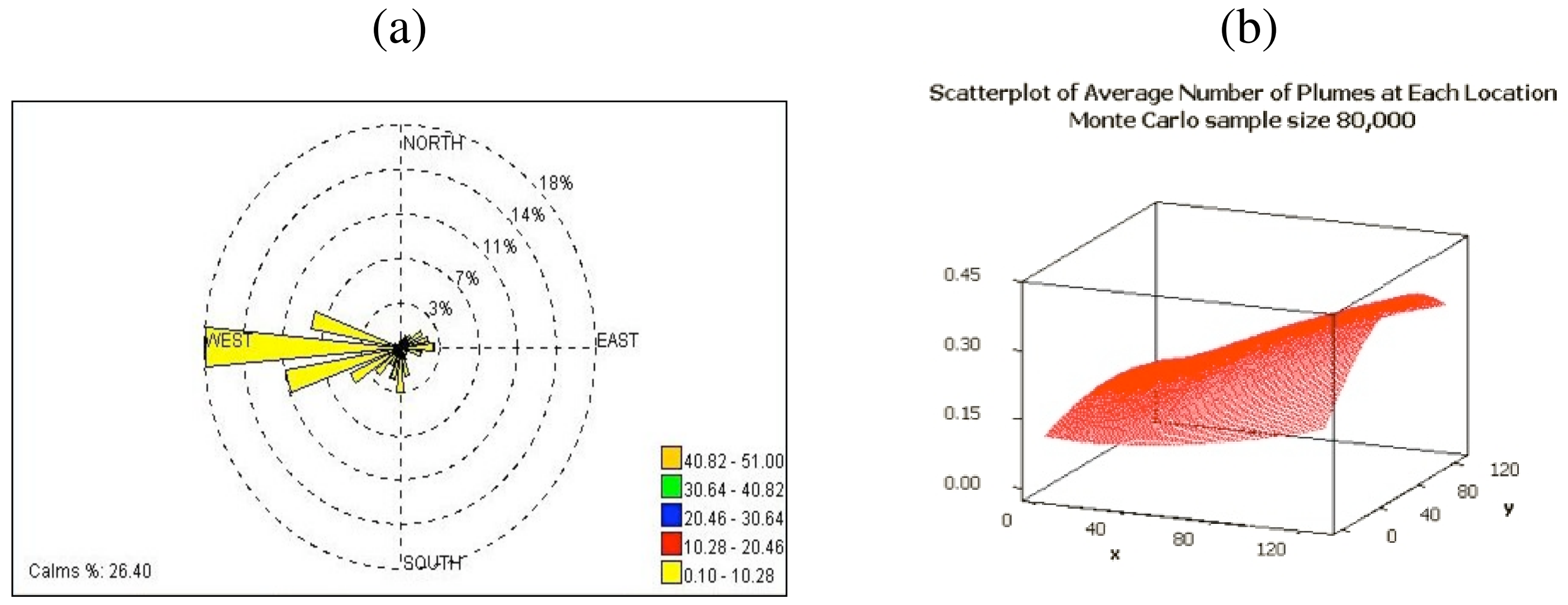
[†]University of Southampton, UK & [‡]Dstl, UK ([⋆]M.Feddag@soton.ac.uk)

Motivating example: dispersion model

The work is motivated by dispersion computer models, which typically have the following features:

1. The input variables are usually of two types (meteorological and source), and can be quantitative or qualitative.
2. There is substantial prior information about the distribution of the input variables from, for example, empirical observations (meteorological) or expert prior knowledge (source) – see Figure 1 (a).
3. These prior distributions are not usually independent, either within type (for example, wind direction and speed is defined via a wind rose) or between type (wind direction and source location).
4. The distributions define a joint probability density (or weight function) on the design region, which is likely to have substantial areas of low weight. It is therefore undesirable to observe responses from input combinations from such areas.

Although dispersion models are generally quick to evaluate, when used routinely in, for example, sensor placement algorithms or as part of other optimization functions, there is a need to reduce the number of code evaluations through carefully designed computer experiments (Santner et al., 2003). Each run of the dispersion model, for given values of the meteorological and source variables, produces a two-dimensional function or plume (doseage across a geographical plane). Current practice is to Monte Carlo sample from the prior distributions for these variables to produce an "average" doseage response surface – see Figure 1 (b).


(a)

(b)

Figure 1: (a) The wind rose representing prior information used to generate the Monte Carlo sample shown in Figure 1 (b). The wind speeds are taken in knots. (b) The doseage surface obtained from averaging the number of plumes seen at each location (x,y) over 80,000 generated plumes.

Weighted space-filling criterion

Consider k_1 quantitative variables $x_1, \ldots, x_{k_1}$ and k_2 qualitative variables $x_{k_1+1}, \ldots, x_{k_1+k_2}$, with qualitative variable x_j having m_j levels denoted by $\mathcal{M}_j = \{1, \ldots, m_j\}$ $(j = k_1+1, \ldots, k_1+k_2)$. Following Qian et al. (2008), we define the distance between two points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X} = \mathcal{R} \times \prod_j \mathcal{M}_j$, where $\mathcal{R} \subset \mathbb{R}^{k_1}$, as

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{k_1} (x_i - y_i)^2 + \alpha \sum_{j=k_1+1}^{k_1+k_2} \mathbf{I}[x_j \neq y_j]}\,, \qquad (1)$$

where $\mathbf{I}[r \neq s]$ is the indicator function that takes the value 1 if $r \neq s$ and 0 otherwise; (1) is a weighted sum (with respect to $\alpha > 0$) of the L_2 distance for quantitative variables and the 0-1 distance for qualitative variables. To account for known dependencies between two variables, we redefine the distance metric $d(\boldsymbol{x}, \boldsymbol{y})$ to include the weight function $w(\boldsymbol{y})$, derived from a joint distribution function across the $k_1 + k_2$ variables:

$$d^\star(\boldsymbol{x}, \boldsymbol{y}) = w(\boldsymbol{y}) d(\boldsymbol{x}, \boldsymbol{y}).$$

The distance between a design $d = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\} \in \mathcal{X}^n$ and a point $\boldsymbol{y} \in \mathcal{X}$ is then defined as

$$D^\star(d, \boldsymbol{y}) = \min_{x \in d} d^\star(\boldsymbol{x}, \boldsymbol{y}).$$

We interpret the weight function as a measure of interest in observing a response at the point $\boldsymbol{y}$. If $w(\boldsymbol{y}) = 0$, all designs are considered to be arbitrary close to y, and the addition of y will not enhance the design's space filling properties. We define the $U^\star$-optimality criterion, and seek a design that minimizes

$$\phi(d) = \int_{\mathcal{X}} D^\star(d, y)\, \mathrm{d}W(\boldsymbol{y})\,, \qquad (2)$$

where $W(\cdot)$ is the distribution function corresponding to $w(\cdot)$. For $w(\boldsymbol{y}) = 1\, \forall\, \boldsymbol{y} \in \mathcal{X}$, (2) reduces to the standard U-optimality space-filling criterion (PROC Optex, SAS, 1995).

Illustrative example

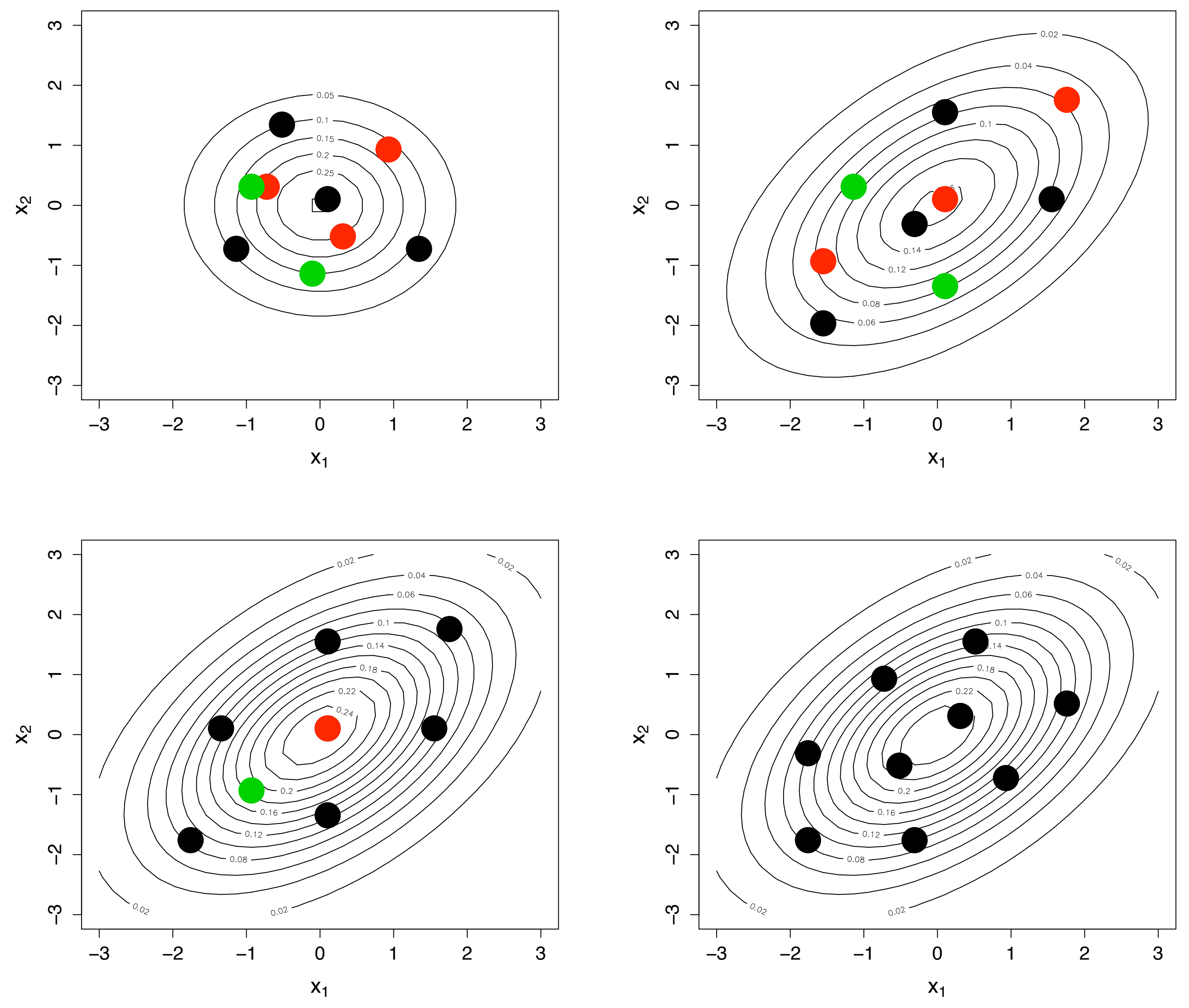


Figure 2: Contour plots for the variables x_1, x_2 at level 1 of x_3, with parameters Σ_1, p_1, α_1 (top left), Σ_2, p_1, α_1 (top right), Σ_2, p_2, α_1 (bottom right) and Σ_2, p_2, α_2 (bottom left).

Consider three input variables: x_1, x_2 quantitative variables defined on $[-3, 3]$ and x_3 a categorical variable with three levels $0, 1, 2$, and a weight function defined as a product of the joint density of (x_1, x_2) from a $N(\mathbf{0}, \Sigma)$ distribution and discrete probabilities $p = (p_1, p_2, p_3)$ for the levels of x_3:

$$g(\boldsymbol{x}, \Sigma, p) = \frac{p}{2\pi \mid \Sigma \mid^{1/2}} \exp\left[-\frac{1}{2}\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}\right]$$

Figure 2 shows weighted space-filling designs for this example using the following parameters: $\Sigma = I_2$ (denoted Σ_1) and $I_2 + J_2$ (Σ_2), $p = (1/3, 1/3, 1/3)$ (p_1) and $(1/2, 1/4, 1/4)$ (p_2), and $\alpha = 1$ (α_1) and $1/2$ (α_2).

Application to the dispersion model

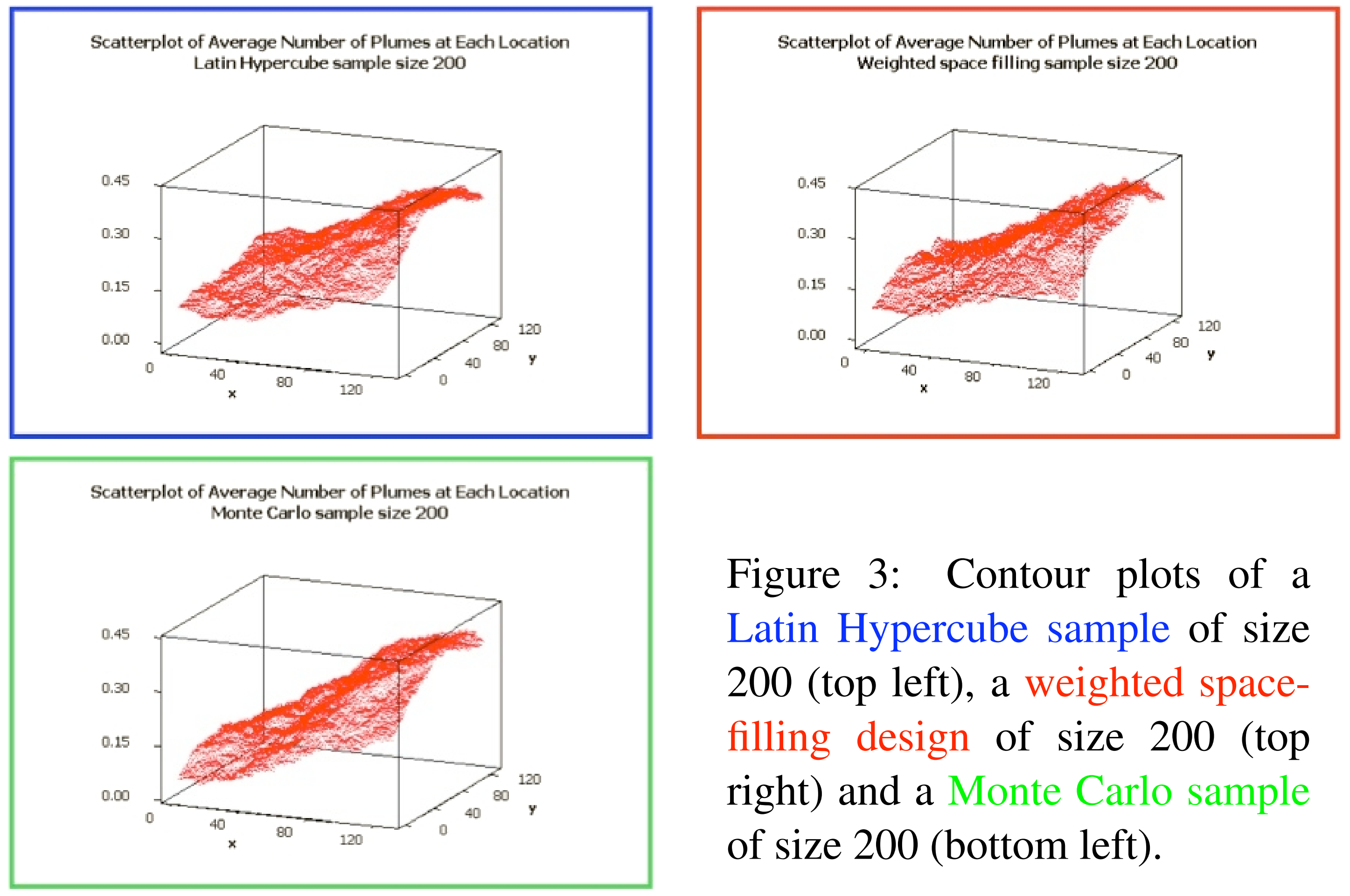


Figure 3: Contour plots of a Latin Hypercube sample of size 200 (top left), a weighted space-filling design of size 200 (top right) and a Monte Carlo sample of size 200 (bottom left).

Table 1: Results from comparison with the 'true' surface.

Squared Error	Monte Carlo	Latin hypercube	Weighted Space-filling
Mean	0.0070	0.0006	0.0004
St. Dev.	0.0085	0.0007	0.0007
Maximum	0.0410	0.0050	0.0054

A weighted space filling design for 7 quantitative meteorological and source variables were found with, for example, non-uniform and dependent prior distributions for wind speed and direction; Figure 1 (a). A comparison with a Latin Hypercube design (McKay et al., 1979; Iman and Conover, 1982) shows similar performance (Figure 3), with both methods producing surfaces that resemble the true surface (from 80,000 Monte Carlo samples, Figure 1 (b)). Figure 3 also shows the results from a Monte Carlo sample of 200 runs that underperforms in comparison to both designs. Table 1 further illustrates this difference, in terms of the mean squared error.

Acknowledgements

This work was supported by DTRA Basic Research Grant HDTRA-08-0048 and the Research Acquisitions Office.

References

Iman, R.L. and Conover, W.J. (1982). A distribution-free approach to inducing rank correlation among input variables. Communications in Statistics - Simulation and Computation, 11, 311-334.

McKay, M.D., Beckman, R.J. and Conover, W.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21, 239-245.

Qian, P.Z.G, Wu, H. and Wu, C.F.J. (2008). Gaussian Process Models for Computer Experiments with Qualitative and Quantitative Factors. Technometrics, 50, 383-396.

Santner, T.J., Williams, B.J and Notz, W.I. (2003). The Design and Analysis of Computer Experiments. Springer, New York.

SAS (1995). SAS QC Software, Vol 1: Useage and Reference. SAS Institute, Inc., Cary, NC.