Annexure C: Comparison of Two Response Surface Methodologies

Network Modeling Analysis in Health Informatics and Bioinformatics
Application of Central Composite Face-Centred (CCF) and Box Behnken Designs (BBD) to study the effect of demographic characteristics on HIV risk in South Africa.

WILBERT SIBANDA
DISTM NU PRE-CLINICAL PLATFORM
FACULTY OF HEALTH SCIENCES
POSTGRADUATE CAMPUSS
NORTH WEST UNIVERSITY
P.O. Box 20406
VANDERBURGH 2000
SOUTH AFRICA
Email: Wilbert.sibanda@nwu.ac.za

PHILIP PRETORIUS
SCHOOL OF INFORMATION TECHNOLOGY
VAAL TRINITY CAMPUS
NORTH WEST UNIVERSITY
Van Blerk Blvd, Vanderburg
SOUTH AFRICA
Email: Philip.pretorius@nwu.ac.za

Abstract- In this study, a Central composite face-centered and Box-Behnken design were employed to study the main and interaction effects of demographic characteristics on the risk of HIV in South Africa. The demographic characteristics studied for each pregnant mother attending an antenatal clinic in South Africa were mother’s age, partner’s age (father’s age), mother’s level of education and parity.

The central composite face-centered and Box-Behnken designs showed that the risk of acquiring HIV status for an antenatal clinic attender was highly sensitive to changes in pregnant women’s age and her educational level, using the 2005 South African annual antenatal HIV and syphilis seroprevalence data.

Individually the age of the pregnant woman’s partner and her parity had no significant effect on the HIV risk. However, the latter two demographic characteristics exhibited significant effects on the HIV risk in two way interactions with other demographic characteristics.

Using HIV as the optimization objective, the following summary statistics were obtained: R² = 0.99 (Central composite face-centered design) and R² = 0.99 (Box-Behnken design). The two-factor interactions model F-values for the central composite and Box-Behnken designs were 5.99 and 68.29 respectively. These F-values for the 2-factor interactions were significant with only a 0.12% and 0.01% chance that these model values were a result of noise respectively. Adeq. precision values of 8.84 and 31.33 for the central composite face-centered and Box-Behnken respectively, suggested that these 2 factor interactions models could be used to navigate the design spaces.

Finally, the main effects and interactions plots illustrated that the HIV risk increased with the age of the pregnant woman.

Keywords- Response surface design, central composite design, face centered, demographic characteristics, seroprevalence data.

1. INTRODUCTION

In South Africa, the annual antenatal HIV survey is the only existing national surveillance activity for determining HIV prevalence and it is therefore a vitally important tool to track the geographic and temporal trends of the epidemic [3].

Raw antenatal clinic data contains the following demographic characteristics for each pregnant woman; age (herein called mother’s age), population group (race), level of education (herein called education), gravidity (number of pregnancies), parity (number of children born), partner’s age (herein called father’s age), name of clinic, HIV and syphilis results [3].

This research paper explores the application of response surface methodologies (RSM) to study the intricate relationships between antenatal data demographic characteristics and one response variable (HIV risk). An RSM is a collection of mathematical and statistical techniques used for modelling and analysis of problems in which a response of interest is influenced by several variables and the objective is to optimize this response [4].

The RSM methodologies used in this research were the central composite face-centered and the Box-Behnken designs, devised by George E.P. Box and Donald Behnken. This study follows up on our previous work [3] where a two level fractional factorial design was used to develop a ranked list of important through unimportant demographic characteristics affecting the HIV status of pregnant mothers in South Africa. Further work was conducted to investigate the possible effect of interaction of demographic characteristics on the HIV risk using a central composite face-centered design [9]. The present study is attempting to prove that the results obtained using the central composite face-centered were not design specific.
The two-level fractional factorial design ranked the relative contributions of the demographic characteristics towards influencing the risk of acquiring the HIV infection using a Lenth’s plot, as shown in Fig. 1.

According to the Lenth plot, the further from the diagonal the demographic characteristic is, the more influential it is on the risk of acquiring HIV infection. The results from the Lenth’s plot can be summarized as follows:

- mother’s age > level of education > parity > father’s age > gravidity > syphilis.

The summary statistics from the two level fractional factorial design are outlined in Table 1.

<table>
<thead>
<tr>
<th>Summary</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.84</td>
</tr>
<tr>
<td>$R^2$ adjusted</td>
<td>0.76</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.18</td>
</tr>
<tr>
<td>PRESS</td>
<td>0.52</td>
</tr>
<tr>
<td>$R^2$ for Prediction</td>
<td>0.34</td>
</tr>
<tr>
<td>First Order Autocorrelation</td>
<td>-0.74</td>
</tr>
<tr>
<td>Collinearity</td>
<td>0.83</td>
</tr>
<tr>
<td>Coefficient of Variation</td>
<td>52.17</td>
</tr>
<tr>
<td>Precision Index</td>
<td>9.96</td>
</tr>
</tbody>
</table>

As shown in Table 1, the adjusted $R^2$ (coefficient of determination) value for the fitted model was 0.76. The $R^2$ value provides the proportion of variability in a data set that is accounted for by the statistical model and it provides a measure of how well future outcomes are likely to be predicted by the model. In other words, the $R^2$ value provides us with information about the goodness of fit of our model. Judging from the size of the adjusted $R^2$ value of the fractional factorial model, this suggested that perhaps an employment of a response surface model would assist in elucidating the possible effect of interaction of demographic characteristics on the regression model. This belief was further substantiated by the low value of the F-statistic ($F=10.64$). The F value indicates the overall significance of the regression model and is thus used to decide whether the model as a whole has statistically significant predictive capability.

2. LITERATURE REVIEW

2.1. Response Surface Methodology (RSM)

An RSM is a collection of statistical and mathematical methods that are useful for modelling and analyzing design. RSM experiments are designed to allow for the estimation of interaction and quadratic effects, and thus providing an idea about the local shape of the response surface being investigated.

Linear terms alone produce models with response surfaces that are hyperplanes. The addition of interaction terms allows for warping of the hyperplane. Squared terms produce the simplest models in which the response surface has a maximum or minimum, and so an optimal response.

RSM comprises of fundamentally three techniques [5], namely:

i. Statistical experimental design
ii. Regression modelling
iii. Optimization
2.1.2. Central composite face-centered design

A central composite face-centered design (Fig. 3) is an example of an RSM that is widely used for fitting a second-order response surface [6]. This design involves the use of a two-level factorial combined with axial points, factorial points, and center runs. The factorial points represent variance-optimal design for a first order and center runs provide information about the existence of curvature in the system [7]. The addition of axial points allows for the efficient estimation of the pure quadratic terms. In that regard, the central composite face-centered design is useful for experiments when there is need to fit a second order response surface.

2.1.3. Box-Behnken design

The Box-Behnken design (Fig. 3) is an independent quadratic design, that does not contain an embedded factorial or fractional factorial design. The Box-Behnken design is characterized by treatment combinations at the midpoints of edges of the experimental space and the centre. These designs are rotatable and require 3 levels of each factor. The designs have limited capability for orthogonal blocking compared to the central composite designs [11].

3. EXPERIMENTAL METHODOLOGY

3.1. Sources of Data

The 2007 antenatal HIV seroprevalence data was obtained from the National Department of Health of South Africa [1]. The data consisted of about 12 000 subjects that attended antenatal clinics for the first time across the nine provinces of South Africa in 2007.

3.2. Research Tools

This research utilized the following research tools:

a) Design Expert V8 Software (StatEase Inc, 2011)

b) SAS 9.3, an integrated system of software products provided by SAS Institute Inc.

c) Essential Regression and Experimental Design, version 2.2 (Gibbons, PA)


3.3. Sampling Procedure

The pre-processing procedure involved the complete randomization of the data in order to reduce bias in the design of experiment.

3.4. Missing Data

The total database consisted of 31 808 cases and of these 21 646 (32%) were found to be complete. However, 10 162 (32%) cases were incomplete and thus discarded.

3.5. Variables

The variables used in the study were mother’s age, partner’s age (herein called father’s age), education level, parity and HIV status. The integer value representing level of education stands for the highest grade successfully completed, with 13 representing tertiary education. Parity represents the number of times the individual has given birth. Parity is important as it shows the reproductive activity as well as reproductive health state of the women.
The HIV status is binary coded; a 1 represents positive status, while a 0 represents a negative status.

3.6. Experimental Design

In this study, the aim was to use a central composite face-centered and a Box-Behnken design to study the individual and interaction effects of demographic characteristics on the HIV risk amongst pregnant women using seroprevalence data. Designs with four factors and one response variable were developed. Based on sparsity-of-effects principle, two factor-interaction design models were used, with 21 runs and no blocks.

3.6.1. Design Matrix Evaluation

3.6.1.1. Degrees of freedom

Design matrix evaluation showed that there were no aliases for the 2 factor interactions and the degrees of freedom (df) for the matrix are shown in Table 2. As a rule of thumb, a minimum of 3 lack-of-fit df and 4 pure error df ensure a valid lack of fit test. Fewer df tend to lead to a test that may not detect lack of fit [8].

<table>
<thead>
<tr>
<th>Degrees of Freedom for the Design Matrices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central composite design</td>
</tr>
<tr>
<td>----------------------------</td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Residuals</td>
</tr>
<tr>
<td>Lack of Fit</td>
</tr>
<tr>
<td>Pure Error</td>
</tr>
<tr>
<td>Cor total</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

3.6.1.2. Standard Errors

The standard errors of the central composite face-centered and Box-Behnken designs are shown in Fig. 4 and 5 respectively. The Box-Behnken design has larger standard errors at the edges of the design space compared to the CCF and this could be attributed to the architecture of the designs. The Box-Behnken design is not capable of estimating the response parameter at the edges of the experimental space. It is therefore advisable to work well within the design margins to achieve a greater degree of accuracy.

3.6.1.3. Fraction of Design Space (FDS)

FDS curve (Fig. 6) is the percentage of the design space volume containing a given standard error of prediction or less. Flatter FDS curve means that the overall prediction error is constant.

In general the larger the standard error of prediction, the less likely the results can be repeated, and the less likely that a significant effect will be detected.
models such as linear model, linear model with interaction terms, linear models with quadratic terms or for full quadratic model. VDGs were developed by Giovannetti-Jensen and Myers in 1999 [12].

Comparison of the VDG graphs (Fig. 8) of the central composite face-centered and Box-Behnken designs illustrates the following points:

- The graphs for the Box-Behnken design are superimposed on each other.
- The Box-Behnken design shows a greater deviation from rotatability than the central composite face-centered design.
- Variance is close for the two designs near the centre.
- Variances of the central composite face-centered and Box-Behnken designs between radii 0.5 and 2 appear to be better for the central composite face-centered design.
- Maximum variance for the central composite face-centered and the Box-Behnken designs are 60 and 160 respectively.

FDS plots for the central composite and Box-Behnken designs (Figs. 7 and 8), indicate that 73% of the central composite face-centered design space is precise enough to predict the mean within ±90, compared to 31% of the Box Behnken.

3.6.1.4. Variance Dispersion Graphs (VDGs)

Variance dispersion graphs have recently become popular in aiding the choice of a response surface design [12]. Furthermore, variance dispersion graphs can be used to compare the performance of multiple designs for a specific
4. RESULTS

4.1. Model Fit Statistics

4.1.1. Sequential model sum of squares

This technique shows the effect of increasing terms to the complexity of the total model.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>F-value</th>
<th>P-value (P&gt;0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CCF</td>
<td>BBD</td>
<td>CCF</td>
</tr>
<tr>
<td>Linear vs. Mean</td>
<td>0.041</td>
<td>0.097</td>
<td>1.51</td>
</tr>
<tr>
<td>2FI vs. Linear</td>
<td>0.081</td>
<td>0.04</td>
<td>10.59</td>
</tr>
</tbody>
</table>

The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming the null hypothesis is true. At a p-value less than 0.05, the null hypothesis is rejected for both the central composite and Box-Behunken designs and it is concluded that the two-factor interaction models are more significant than linear main effects models.

4.1.2. Model Summary Statistics

The $R^2$ statistics of the linear models were considerably lower than those of the two-factor interactions (2FI) models. The two-factor interactions model have the lowest standard deviation and a high $R^2$. The two design methodologies provided exactly the same $R^2$ values. As explained before, high $R^2$ values imply that a large proportion of the variation in the observed values is explained by the model.

Adeq precision is used to measure the signal to noise ratio and a ratio greater than 4 is desirable indicating model can be used to navigate the design space. The central composite face-centered and Box-Behunken designs provided adeq precision values of 8.84 and 31.33 respectively indicating an adequate signal. Therefore these models can be used to navigate the design space.

4.2. ANOVA for two-factor interactions response surface

The ANOVA table (Table 6) for both CCF and BBD confirms the adequacy of the 2FI model. The model F-value of 7.99 for the CCF is significant and only a 0.12% chance that this model F-value could be due to noise. In addition, the model F-value of 88.29 for the BBD is also significant with only a 0.01% chance of this value being due to noise.
The two designs confirmed that the mother’s age had the greatest effect on the risk of acquiring HIV infection. The mother’s educational level was the second most important individual factor. Furthermore, the interaction of the mother’s age with her partner’s age and educational level were found to be significant towards influencing the risk of acquiring HIV infection.

**TABLE 6: ANOVA RESULTS**

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum Of Squares</th>
<th>F-value</th>
<th>P-value (P&gt;0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CCF</td>
<td>BBD</td>
<td>CCF</td>
</tr>
<tr>
<td>Model</td>
<td>0.12</td>
<td>0.014</td>
<td>7.09</td>
</tr>
<tr>
<td>Mother’s age</td>
<td>0.035</td>
<td>0.054</td>
<td>18.2</td>
</tr>
<tr>
<td>Father’s age</td>
<td>0.004</td>
<td>0.000</td>
<td>0.02</td>
</tr>
<tr>
<td>Education</td>
<td>0.012</td>
<td>0.005</td>
<td>6.09</td>
</tr>
<tr>
<td>Parity</td>
<td>0.000</td>
<td>0.001</td>
<td>0.13</td>
</tr>
<tr>
<td>Mot*Edu</td>
<td>0.015</td>
<td>0.023</td>
<td>7.07</td>
</tr>
<tr>
<td>Mot*Par</td>
<td>0.057</td>
<td>0.017</td>
<td>29.7</td>
</tr>
<tr>
<td>Fat*Par</td>
<td>0.005</td>
<td>0.001</td>
<td>2.35</td>
</tr>
<tr>
<td>Edu*Par</td>
<td>0.000</td>
<td>0.000</td>
<td>22.6</td>
</tr>
</tbody>
</table>

5. MODEL ADEQUACY CHECKING

Model adequacy checking is conducted to verify whether the fitted model provides an adequate approximation to the true system and to verify that none of the least squares regression assumptions are violated. Extrapolation and optimization of a fitted response surface will give misleading results unless the model is an adequate fit.

There are many statistical tools for model validation, but the primary tool for most process modeling applications is graphical residual analysis.

5.1. Residual Analysis

5.1.1. Normality Probability Plot of Residuals

A normal probability plot of residuals can be used to check the normality assumption. If the residuals plot approximates a straight line, then the normality assumption is satisfied. The normal plot of residuals also evaluates whether there are outliers in the dataset. If all the points lie on the diagonal, this implies that the residuals constitute normally distributed noise as shown by the plot for the central composite face-centered design. However, the normal probability plot (for residuals for the Box-Behnken design was slightly curved, perhaps indicative of non-modelled quadratic relations or insufficient transformations.

![Normal Plot of Residuals](image)

**Fig. 9: Normal plot of residuals for the CCF Design**
5.1.2. Plot of Residuals vs Predicted Response

The residuals should scatter randomly suggesting that the variance of the original observations is constant for all values of the response. However, if the variance of the response depends on the mean level of the response, the shape of the plot tends to be funnel-shaped, suggesting a need for a transformation of the response variable.

5.1.3. Plot of Residuals vs Observation order

The plots of residuals vs observation order for central composite face-centered and Box-Behnken designs are shown in Fig. 13 and 14. The random patterns on these plots indicate model adequacy, suggesting no need for transformation to stabilize the situation.
5.2. Influence Diagnosis

Parameter estimates or predictions may depend more on the influential subset than on the majority of the data. It is therefore important to locate these influential points and assess their impact on the model. One approach for exploring influential points is to use their leverage.

5.2.1. Leverage Points

This is a measure of the disposition of points on the x-space. As stated above, some observations tend to have disproportionate leverage on the parameter estimates, the predicted values and the summary statistics [5].

The leverage of points patterns were similar for both the central composite face-centered and Box-Behunen designs.

6. FINAL EQUATIONS OF THE RESPONSE SURFACE MODELS

<table>
<thead>
<tr>
<th>CCF</th>
<th>BBD</th>
<th>Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV=</td>
<td>HIV=</td>
<td></td>
</tr>
<tr>
<td>+0.33</td>
<td>+0.32</td>
<td></td>
</tr>
<tr>
<td>+0.15*</td>
<td>+0.18*</td>
<td>Mother’s age</td>
</tr>
<tr>
<td>-0.002*</td>
<td>-0.02*</td>
<td>Father’s age</td>
</tr>
<tr>
<td>+0.09*</td>
<td>-0.07*</td>
<td>Education</td>
</tr>
<tr>
<td>+0.005*</td>
<td>+0.02*</td>
<td>Parity</td>
</tr>
<tr>
<td>+0.04*</td>
<td>+0.13*</td>
<td>Mother's age *Father's age</td>
</tr>
<tr>
<td>-0.11*</td>
<td>-0.10*</td>
<td>Mother’s age *Education</td>
</tr>
<tr>
<td>-0.02*</td>
<td>+0.05*</td>
<td>Mother’s age *Parity</td>
</tr>
<tr>
<td>-0.02*</td>
<td>+0.02*</td>
<td>Father's age *Parity</td>
</tr>
</tbody>
</table>
A main effects plot is a plot of the means of the response variable for each level of a factor, allowing for the determination of which main effects are important.

For both the central composite face-centered and the Box-Behnken designs, HIV risk increases steeply as the mother’s age and her education increase from the low level (-1) to the middle level (0). Thereafter, the HIV risk decreases gradually for the two demographic characteristics.

6.2. Interaction plot

This research assumes sparsity-of-effects principle that states that a system is usually dominated by main effects and low-order interactions. Thus it is most likely that main effects and two-factor interactions are the most significant responses in an experimental design. In other words, higher order interactions such as three factor interactions are rare. This is sometimes referred to as the hierarchical ordering principle. Therefore, based on sparsity-of-effect principle, both the central composite face-centered and the Box-Behnken designs reflected that the interaction of the mother’s age with the other demographic characteristics had a significant effect on the HIV risk of pregnant mothers.
7. DISCUSSION

This research paper was aimed at comparing the applicability of a central composite face-centered (CCF) and Box-Behnken designs (BBB) in determining the differential effects of an individual’s demographic characteristics on the risk of acquiring an HIV infection. The two response surface methodologies produced the same results. Based on the obtained results, the two response surface designs were found to be suitable for studying the involvement of demographic characteristics in predicting the inherent risk of acquiring HIV infection. Furthermore, the two-factor interaction (2FI) polynomial functions for the mother’s age, father’s age, education and parity obtained using StatEase Design Expert and Matlab softwares were found to be statistically significant compared to linear main-effect models.

8. CONCLUSION

The central composite face-centered and the Box-Behnken designs confirmed the results obtained by the fractional factorial design [3], that mother’s age had the greatest single effect on the risk of acquiring an HIV infection amongst pregnant women in South Africa. In addition, the two response surface methodologies demonstrated that interaction of these main effects also contributed to the risk of an HIV infection. This information can be of tremendous value to policy makers involved in formulating strategies to curb the spread of the HIV/AIDS pandemic in South Africa.

It is paramount to fully understand the drivers of the HIV pandemic in order to arrest its spread within a community. An appreciation of the fact that the mother’s age seems to be directly related to increased chances of being HIV positive, between the ages of 15 and 30 years old, provides an opportunity to target these susceptible age groups in order to raise awareness of safer sexual practices to prevent the spread of the HIV epidemic. It was also observed that sexual involvement between very young mothers and older men predisposed the women to higher rates of HIV infection. Again this information, can be used to advise young women to be sexual involved with men within their age groups to reduce chances of HIV infection.

As expected, increased education levels amongst antenatal clinic attendees, reduced the risk of acquiring the infection. Interestingly, parity on its own was found to be non-significant, and this can be explained by the fact that women with more children are generally the older women who have been found to be less at risk of acquiring the infection.

The inability to target spending is one of the most important constraints which governments face in their effort to mitigate the impact of HIV/AIDS. Ideally, the government’s response should be quick, involve long-term and short-term strategies. This response should be based on a thorough understanding of the drivers of the HIV/AIDS epidemic.

9. ACKNOWLEDGEMENTS

Wilbert Sibanda acknowledges doctoral funding from South African Centre for Epidemiological Modelling (SACEMA), Medical Research Council (MRC) and North-West University. Special thanks to Cathrine Toleng Sibanda and the National Department of Health (South Africa) for the antenatal seroprevalence data (2006-2007).

10. REFERENCES


