Conference on Applied Statistics in Agriculture | 2001 - 13th Annual Conference Proceedings

SIMULATION STUDY OF SPATIAL-POISSON DATA ASSESSING INCLUSION OF SPATIAL CORRELATION AND NON-NORMALITY IN THE ANALYSIS

Rebecca A. Hensberry

David B. Marx

Daryl Travnicek

Stephen Kachman

See next page for additional authors

Recommended Citation

Hensberry, Rebecca A.; Marx, David B.; Travnicek, Daryl; and Kachman, Stephen (2001). "SIMULATION STUDY OF SPATIAL-POISSON DATA ASSESSING INCLUSION OF SPATIAL CORRELATION AND NON-NORMALITY IN THE ANALYSIS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1213

Author Information

Rebecca A. Hensberry, David B. Marx, Daryl Travnicek, and Stephen Kachman

SIMULATION STUDY OF SPATIAL-POISSON DATA ASSESSING INCLUSION OF SPATIAL CORRELATION AND NON-NORMALITY IN THE ANALYSIS

Rebecca A. Hensberry, David B. Marx, Daryl Travnicek, and Stephen Kachman
Department of Biometry
University of Nebraska-Lincoln
Lincoln, NE 68583-0712

ABSTRACT

Spatial correlation and non-normality in agricultural, geological, or environmental settings can have a significant effect on the accuracy of the results obtained in the statistical analyses. Generalized linear mixed models, spatial models, and generalized linear models were compared in order to assess how critical the inclusion of non-normality and spatial correlation is to the analysis. Spatially correlated data with a Poisson distribution were generated in a completely randomized design (CRD) with 2 treatments and 18 repetitions. Four analyses: spatial Poisson, non-spatial Poisson, spatial normal, and non-spatial normal, were conducted on the simulated data to compare their power functions. The degree of spatial correlation, size of the mean, the dimension of the plots and difference between the two treatment means were altered to investigate how the ability to detect differences between the treatments is affected. In addition, the range covariance parameter was estimated and compared among the spatial models. Some covariance parameter estimates were under-estimated. The size of the field plot and the treatment means were increased to assess their effects on estimation of the range. The Reduced Maximum Likelihood (REML) covariance parameter estimates were compared to those obtained using Maximum Likelihood (ML) estimates. The analysis that incorporated the spatial correlation of the observations and used ML to estimate the covariance parameters had the highest power and most accurate range parameter estimates.

1. INTRODUCTION

One of the most frequently occurring types of agricultural data is obtained by counting the characteristic of interest, such as the number of insects, the number of weed patches or the number of diseased roots or leaves in each experimental unit. These types of data are generally described with the Poisson distribution (Gomez and Gomez, 1984). Often the experimental units that are closer together are more closely related. For example the level of fertility in the field may decrease as one moves across the field. Two plants beside one another have more similar fertility levels than one plant has with another in an opposing corner of the field plot. The relationship between observations a specified distance apart is known as spatial correlation (Goovaerts, 1997).

There are three ways to characterize the spatial correlation present in the data (Cressie, 1991). Traditional statisticians prefer to examine the amount of covariance between the observations a given distance from one another. Time series analysts may use the correlation between observations taken a given amount of time apart to describe the spatial correlation in time. Spatial statisticians choose to use the semivariance, which is a function of the difference between observations squared. In this simulation study, the semivariance is used to model the spatial correlation present. Several semivariograms are available to model the structure of the spatial correlation present in the field. The spherical covariance structure is most commonly used of the various covariance structures available, as it is the most frequently occurring in nature (Clark, 1979). The three covariance parameters used to characterize the spherical structure are the nugget, range, and sill (Journel and Huijbregts, 1978). The sill, the upper asymptote of the semivariogram, is also characterized as the variance of observations that are independent of one another. The range is the distance at which the observations are essentially independent. The nugget effect is a measure of small-scale spatial variation and a quantification of the amount of measurement error present. In this simulation study, the nugget effect is assumed to be zero.

In order to assess how important the inclusion of spatial correlation and the Poisson distribution are in the analysis of data that are known to possess these characteristics, the type I error rate, power to detect treatment differences, and the accuracy of parameter estimates were compared. The objectives of the simulation study were to simulate spatially correlated Poisson data, compare four analyses that have the distributional assumption of Poisson or normal data and either include or exclude the spatial correlation between the observations. Finally, the results were applied to various field settings, such as rectangular and square plots, large and small amounts of spatial correlation, and a small number versus a large number of average counts.

The first part of the data simulation process was to incorporate the spherical covariance structure (Isaaks and Srivastava, 1989). This was done using SAS's PROC IML. Starting with a normal response variable, the spherical covariance structure is assembled and then multiplied by the vector of observations to obtain the spatially correlated response variable. The response variable is then standardized and assigned a normal probability value. The Poisson count value with the same probability as the standard normal response is then assigned. Essentially, the cumulative distribution functions of the normal and Poisson distributions are matched to convert the normal response to a Poisson response (Moser, 2000).

The experimental design consists of two treatments with eighteen observations per treatment in a completely randomized design. A semivariogram range of 2.2 was used to illustrate how the four analyses performed in a situation where the correlation between observations only stretches a small distance. A range of 10 was also used to model a higher level of spatial correlation between observations. Two treatments were used in each simulation. The mean of one treatment remained constant, while the mean of the second was changed for each set of simulations. The mean of the stationary treatment was chosen to be 15, while the mean of the variable treatment was 15, 16, 17, 18, 19, or 20. A second set of smaller means were also analyzed, with the stationary mean equal to 2, and the variable mean having values of 2, 2.5, 3, 4, 5 or 6. For each scenario 5000 replications were analyzed. In summary, for the set of experiments with a square shape, where the experimental units are organized in a 6x6 grid, there is a set of experiments with high and low levels of spatial correlation, and within each level of spatial correlation there is a set of treatment means that is either large or small (Figure 1).

The four analyses used either included or excluded the spatial correlation and non-normality of the simulated data set. For the analysis that included the spatial correlation and incorporated the distributional assumption of Poisson, the SAS GLIMMIX macro was used with the following SAS (v 8.0) programming statement:

```
%glimmix(data=TRT,procopt=method=ml,maxit=200,
    stmts=%str(
    class TRT;
    model N = TRT/ddfm=kr;
    repeated /sub=intercept type=sp(sph)(lat lng);
    lsmeans TRT;
    error=POISSON);
run;
```

The repeated statement allows for correlated errors with a spatial spherical covariance structure. The experimental unit (sub =) which exhibits correlated errors is denoted by the word "intercept", which suggests that the spatial structure is uniform across the entire field area. The non-spatial Poisson analysis was calculated using SAS PROC GENMOD with the following programming statement:

```
PROC GENMOD;
    CLASS TRT;
    MODEL N = TRT/DIST=POISSON TYPE3;
    LSMEANS TRT;
RUN;
```

The spatial analysis with the distributional assumption of normality was employed using PROC MIXED, with the response variable being the transformed value, or square root, of the count. The following programming statement was employed:

```
PROC MIXED method=ml;
    CLASS TRT;
    MODEL SN=TRT/ddfm=kr;
    REPEATED/SUB=INTERCEPT TYPE=SP(SPH)(LAT LNG);
    LSMEANS TRT;
RUN;
```

As in the GLIMMIX macro, the subject equal to the intercept allows for uniform spatial structure across the entire field. The non-spatial normal analysis was completed using PROC MIXED with the following programming statements:

```
PROC MIXED;
    CLASS TRT;
    MODEL SN=TRT/ddfm=kr;
    LSMEANS TRT;
RUN;
```

As with the spatial normal analysis, the Poisson response variable was transformed by taking the square root of the count. The power curves for four analyses were compared for both the square and rectangular designs. The estimates of the range covariance parameter were examined for accuracy. Diagnostic efforts to investigate problems with range estimates were conducted for larger grid sizes and sets of larger equal treatment means. Initially, the analysis was conducted using Reduced Maximum Likelihood (REML) estimation, and consequentially problems with convergence, range overestimation, and incorrect degrees of freedom arose. The known range values were 2.2 and 10, such that ranges greater than 500 were excessive and considered to be overestimates. Due to the above mentioned problems with REML, the analysis was repeated using Maximum Likelihood (ML) estimation of the covariance parameters in an attempt to obtain more reliable results. The results presented are those from the ML analysis. The percent non-convergence, incorrect degrees of freedom and range overestimates were compared for the two estimation methods.

2. RESULTS AND DISCUSSION

The first portion of the analysis examined is the type I error rate, or percent rejection under the null hypothesis, for both rectangular and square designs (Tables 1-2). Overall the rejection rates are fairly close to the expected rate of 0.05 although for a large range the spatial normal method seems to overestimate the type I error rate. However, when there is a large amount of spatial correlation present, the non-spatial Poisson analysis appears to be under-rejecting for both large and small treatment means. The power curves were compared for the four analyses in the four settings: low spatial correlation and small means, low spatial correlation and large means, high spatial correlation and small means, high spatial correlation and large means (Figures 2-5). In all four settings used to compare the power curves for the square design, the two spatial analyses had power greater than or equal to the two analyses that did not include the spatial correlation in the data set. The distributional assumption therefore appears to be less influential on the power to detect treatment differences than the inclusion of known spatial correlation. The power curves obtained from the square and rectangular analyses were compared to determine if any large differences were present in the power for the two design shapes in the four settings (Figures 6-9). There did not appear to be large discrepancies between power curves for the rectangular and square designs. The range covariance parameters were considered to check accuracy of estimation of the analyses when the means are equal (Table 3) and when the means are extremely unequal (Table 4). The range covariance parameter estimates for equal treatment means were underestimated for the spatial Normal analysis when the known range is 2.2. The spatial Poisson analysis is underestimating the known range of 2.2 more than the spatial Normal. When the known range is 10, both spatial analyses are largely underestimating the range covariance parameter. The results are similar for extreme unequal means. Three possible reasons for range underestimation are the proportion of zero estimates present, the size of the design, and the size of the treatment means. The proportion of zero estimates for the range covariance parameter estimates for equal and unequal means, and high and low spatial correlation settings were checked for the two spatial analyses (Tables 5-6). The proportion of zero estimates was approximately 90% for the small range and approximately 70% for the large range in the spatial Poisson analysis. In the spatial normal analysis, the proportions of zero estimates were approximately 70% and 45% for the small and large ranges, respectively. Based on the proportion of zeroes present, the spatial Poisson analysis was less likely to detect spatial correlation than the spatial normal analysis. The relatively large percentage of zero estimates is likely to be partially responsible for the underestimates of the range covariance parameter. The estimation of the range was examined for larger equal treatment means of 100 and 500, and for larger plot sizes to determine effects on the estimates of the range covariance parameters (Table 7). In the 10x10 grid, as the size of the equal treatment means increases, the estimate of the range approaches the true value of 10 for both spatial analyses. When the size of the grid is increased from 6x6 to 20x20 and 30x30, the spatial normal analysis the estimate of the range improves as well. The spatial Poisson range estimate becomes more accurate as the mean of the treatments increases (10x10 grid) yet does not improve as the size of the field plot increases for means of 15 (Table 7). The percent non-convergence, incorrect degrees of freedom and range overestimates were obtained from all the simulations and compared for REML and ML (Table 8). The percent nonconvergence for the REML analysis was 0.020% compared to 0.006% for the ML analysis. The percent incorrect degrees of freedom for REML and ML are 0.0823% and 0, respectively. The REML estimation method yielded 15-30% overestimates, while the ML estimation method had 0.048% overestimates of the range. The ML estimation is preferable to the REML estimation procedure due to lower rates of nonconvergence, lack of incorrect degrees of freedom as well as fewer overestimates of the range.

3. SUMMARY

The analysis that considered the spatial correlation present in the data set had higher power and more accurate rejection rates in a wide variety of settings. The distributional assumption appears to be less important in terms of the ability to detect treatment differences however; the spatial normal analysis yielded more accurate range estimates overall. The maximum likelihood method of parameter estimation provides more favorable results with lower rates of parameter overestimation, no incorrect degrees of freedom and less nonconvergence. Future research is necessary to determine which analysis is best when there are more than two treatments present, the distribution of the data is neither normal nor Poisson, and when the design is a randomized complete block instead of a completely randomized design. These conclusions are based upon using the SAS GLIMMIX macro for the spatial Poisson analysis. Other computational procedures may perform differently.

REFERENCES

Clark, I. 1979. Practical geostatistics. Applied Science Publishers LTD. London

Cressie, N. 1991. Statistics for spatial data. John Wiley and Sons, Inc. New York.

Goovaerts, P. 1997. Applied geosstatistics series, Oxford University Press. New York.

Gomez, K. A. and A. G. Gomez. 1984. Statistical procedures for agricultural research, 2nd edition. John Wiley and Sons, Inc. New York.

Isaaks, E. H. and R. M. Srivastava. 1989. An introduction to applied geostatistics. Oxford University Press, Inc. New York.

Journel, A. G. and C. Huijbregts. 1978. Mining geostatistics. Academic Press. New York.

Moser, B. 2000. Professor of Statistics, Louisiana State University. Personal communication with D. B. Marx.

Zar, J. H. 1974. Biostatistical Analysis. Prentice-Hall, Inc. Englewood Cliffs, NJ.

Table 1. Percent rejection under the null hypothesis for the square design, α =0.05.

	Small Range		**Large Range**	
	$\mu_A = 2$	$\mu_A = 15$	$\mu_A = 2$	$\mu_A = 15$
Analysis:				
Spatial Poisson	0.0464	0.0434	0.0558	0.0575
Non-spatial Poisson	0.0418	0.0354	0.0150	0.0096
Spatial Normal	0.0586	0.0506	0.0694	0.0640
Non-spatial Normal	0.0398	0.0358	0.0418	0.0368

Table 2. Percent rejection under the null hypothesis for the rectangular design, α =0.05.

	Small Range		**Large Range**	
	$\mu_A = 2$	$\mu_A = 15$	$\mu_A = 2$	$\mu_A = 15$
Analysis:				
Spatial Poisson	0.0546	0.0456	0.0574	0.0586
Non-spatial Poisson	0.0492	0.0360	0.0176	0.0120
Spatial Normal	0.0704	0.0538	0.0724	0.0666
Non-spatial Normal	0.0500	0.0378	0.0436	0.0422

Table 3. Range covariance parameter estimates ± $\sigma / \sqrt{n}$ for equal means.

		Range = 2.2	Range = 10
		$\widehat{\text{Range}}$	$\widehat{\text{Range}}$
	Analysis:		
$\mu_A = 2$	Spatial Poisson	0.33±0.02	1.70±0.04
$\mu_B = 2$	Spatial Normal	1.01±0.03	2.55±0.04
$\mu_A = 15$	Spatial Poisson	0.24±0.01	1.81±0.04
$\mu_B = 15$	Spatial Normal	1.22±0.03	3.09±0.04

Table 4. Range covariance parameter estimates ± $\sigma / \sqrt{n}$ for extreme unequal means.

		Range = 2.2	Range = 10
		$\widehat{\text{Range}}$	$\widehat{\text{Range}}$
	Analysis:		
$\mu_A = 2$	Spatial Poisson	0.33±0.02	1.83±0.04
$\mu_B = 6$	Spatial Normal	1.07±0.03	2.70±0.04
$\mu_A = 15$	Spatial Poisson	0.23±0.01	1.86±0.04
$\mu_B = 20$	Spatial Normal	1.15±0.03	3.11±0.04

Table 5. Proportion of zero range estimates for equal means.

		Range = 2.2	Range = 10
	Analysis:		
$\mu_A = 2$	Spatial Poisson	0.9194	0.6768
$\mu_B = 2$	Spatial Normal	0.7346	0.4756
$\mu_A = 15$	Spatial Poisson	0.9392	0.6764
$\mu_B = 15$	Spatial Normal	0.6794	0.3738

Table 6. Proportion of zero range estimates for extreme unequal means.

		Range = 2.2	Range = 10
	Analysis:		
$\mu_A = 2$	Spatial Poisson	0.9174	0.6582
$\mu_B = 6$	Spatial Normal	0.7216	0.4426
$\mu_A = 15$	Spatial Poisson	0.9422	0.6628
$\mu_B = 20$	Spatial Normal	0.6954	0.3728

Table 7. Range estimates for larger means and larger design dimensions.

		$\mu_A = \mu_B =$ **15**:	$\mu_A = \mu_B =$ **100**:	$\mu_A = \mu_B =$ **500**:
		Range = 10	Range = 10	Range = 10
		$\hat{\text{Range}} \pm \sigma/\sqrt{n}$	$\hat{\text{Range}} \pm \sigma/\sqrt{n}$	$\hat{\text{Range}} \pm \sigma/\sqrt{n}$
Size:	Analysis:			
6x6	Spatial Poisson	1.81±0.04		
	Spatial Normal	3.09±0.04		
10x10	Spatial Poisson	7.11 ± 0.77	8.25 ± 1.27	9.39 ± 2.71
	Spatial Normal	9.33 ± 0.39	8.64 ± 0.55	9.53 ± 0.21
20x20	Spatial Poisson	19.10 ± 0.99		
	Spatial Normal	9.83 ± 0.29		
30x30	Spatial Poisson	31.87 ± 1.22		
	Spatial Normal	9.97 ± 0.08		

Table 8. Nonconvergence, incorrect degrees of freedom, and overestimation rate for the REML versus the ML method of covariance parameter estimation.

*Of 0.823% incorrect degrees of freedom only 3% are from simulations with large means (μ_A=15).

	REML	**ML**
Non-convergence	0.020 %	0.006 %
Incorrect degrees of freedom	0.823 %*	0
Range over-estimates:		
(Range > 500)	15% - 30 %	0.048 %

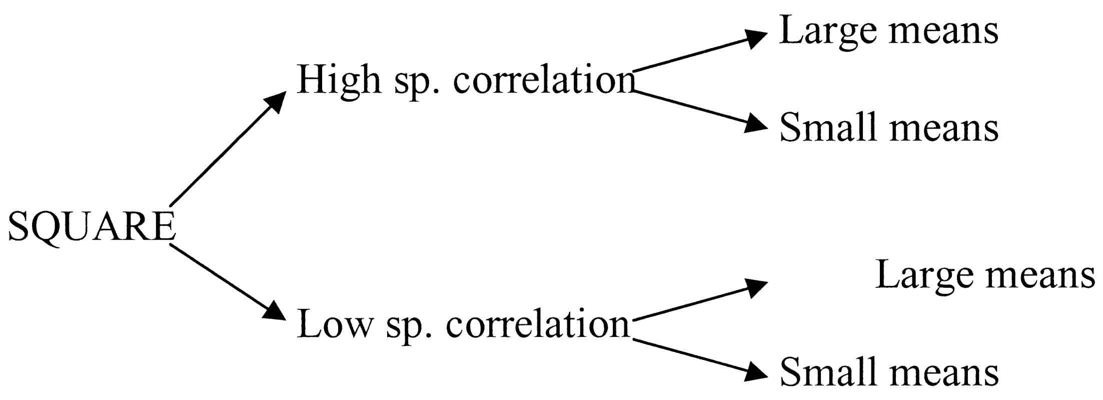


Figure 1. Flow chart showing the components of the simulation procedure.

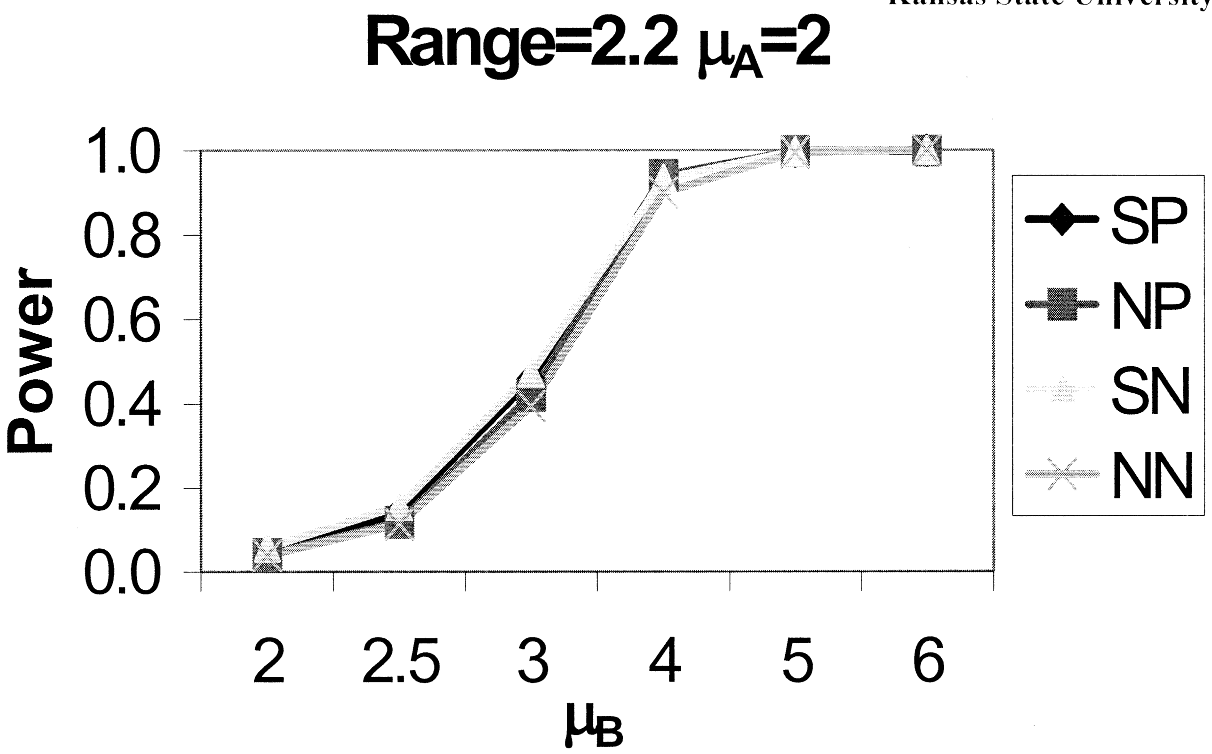
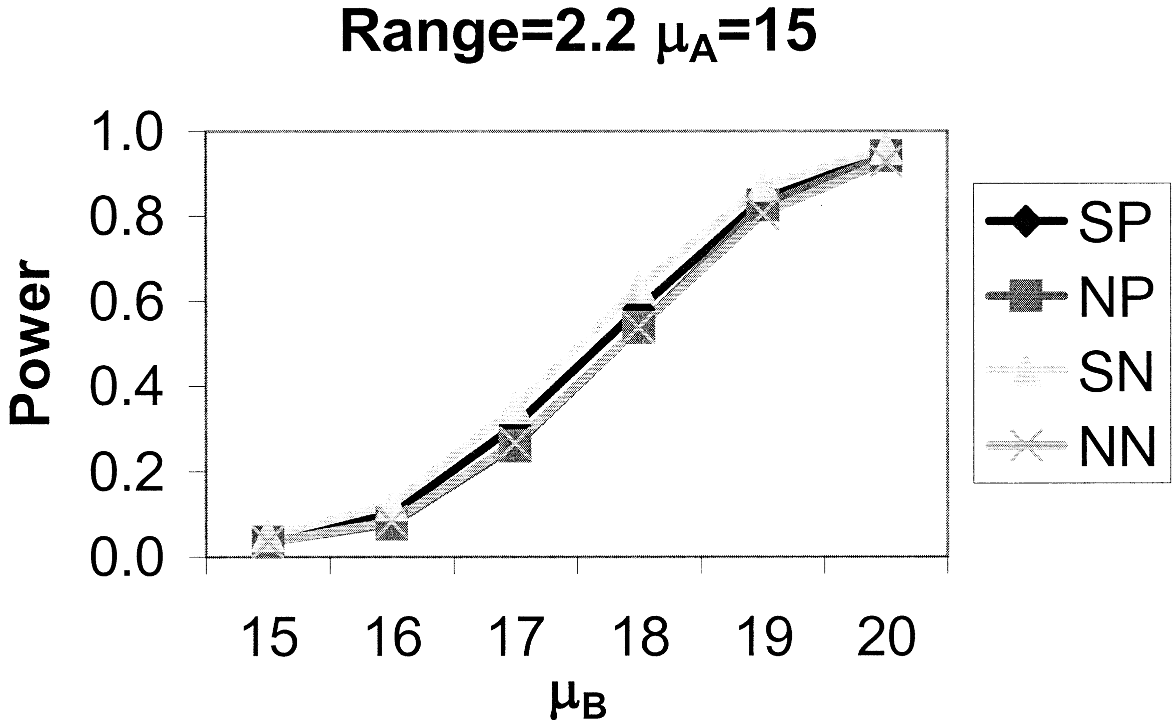


Figure 2. Comparison of power curves for square design with small range and small means



. Figure 3. Comparison of power curves for square design with small range and large means.

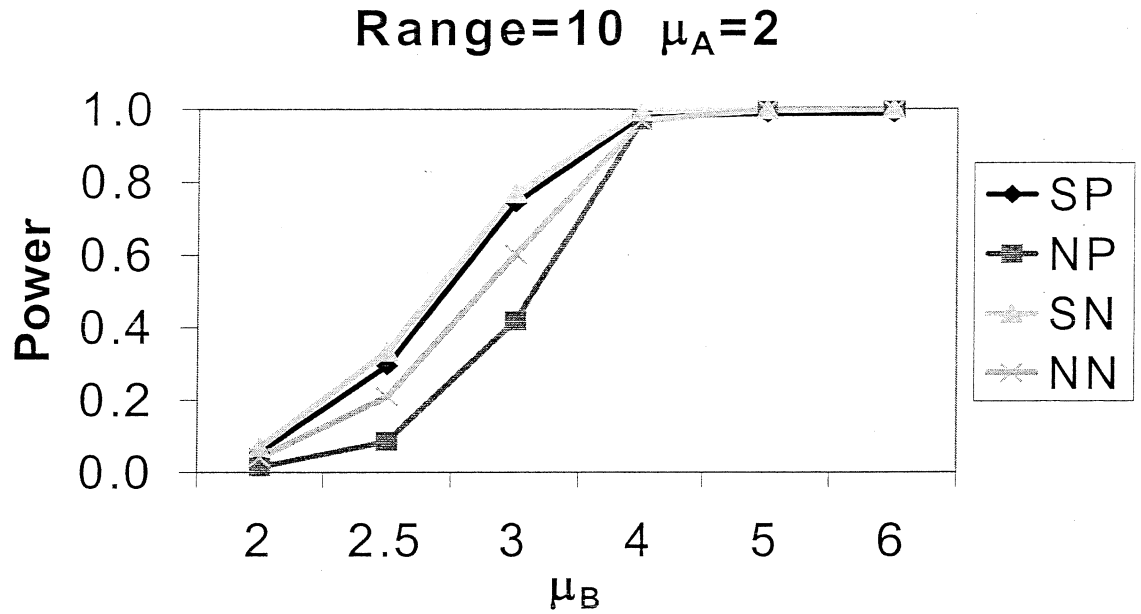


Figure 4. Comparison of power curves for square design with large range and small means.

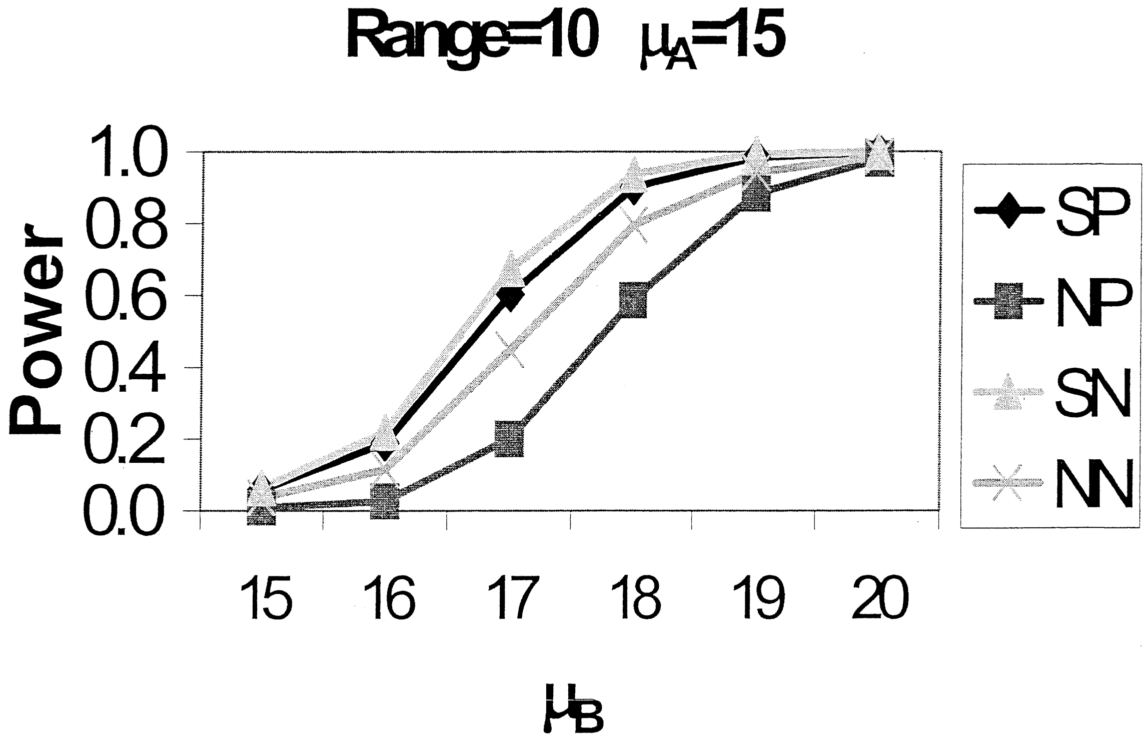


Figure 5. Comparison of power curves for square design with large range and large means

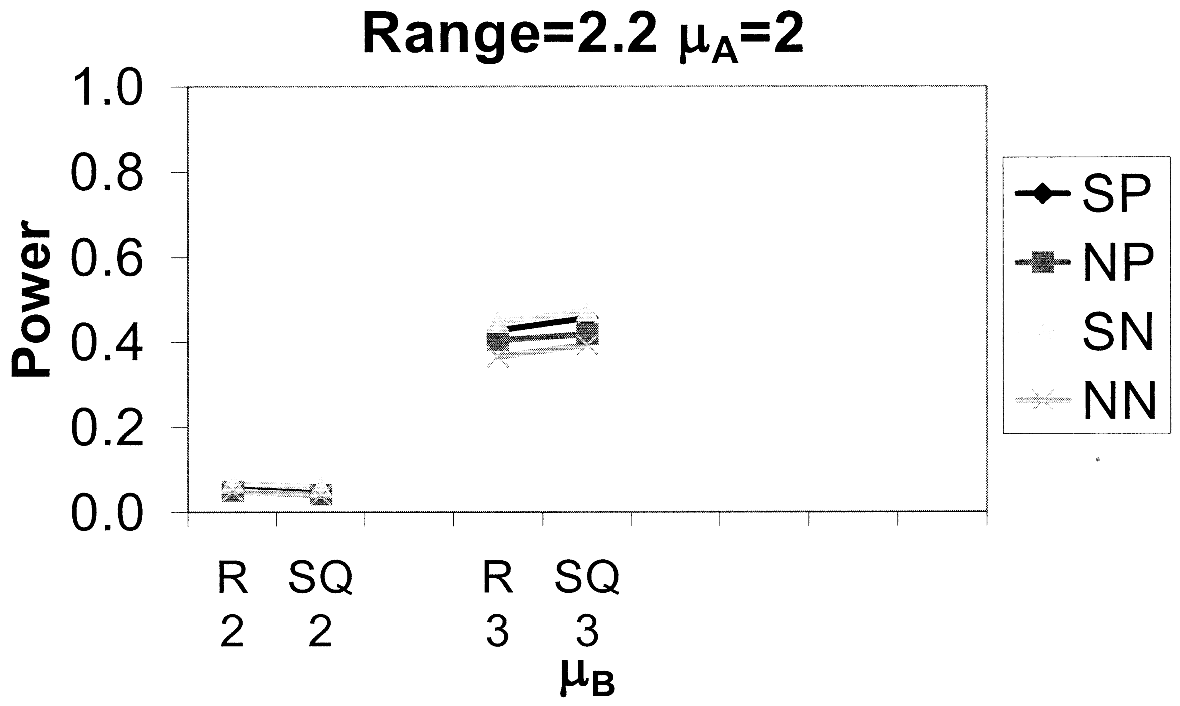


Figure 6. Comparison of power curves of rectangular versus square design for small range and small means.

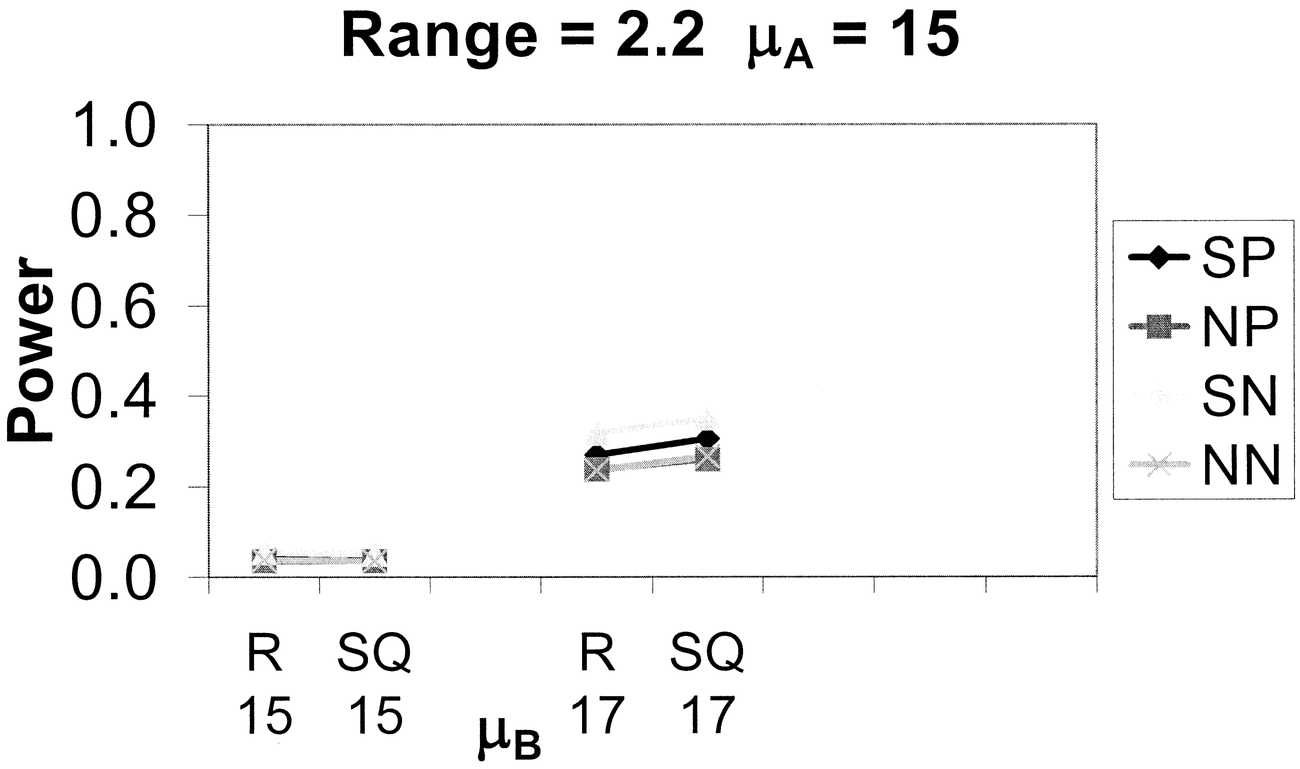


Figure 7. Comparison of power curves of rectangular versus square design for small range and large means.

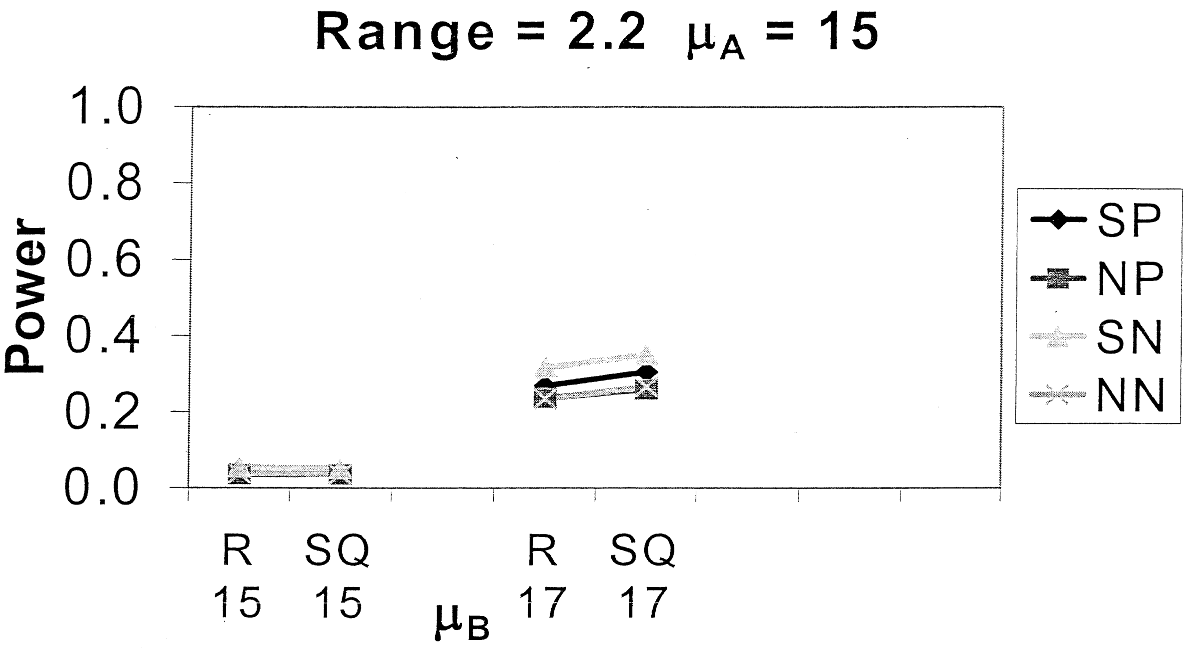


Figure 8. Comparison of power curves of rectangular versus square design for large range and small means.

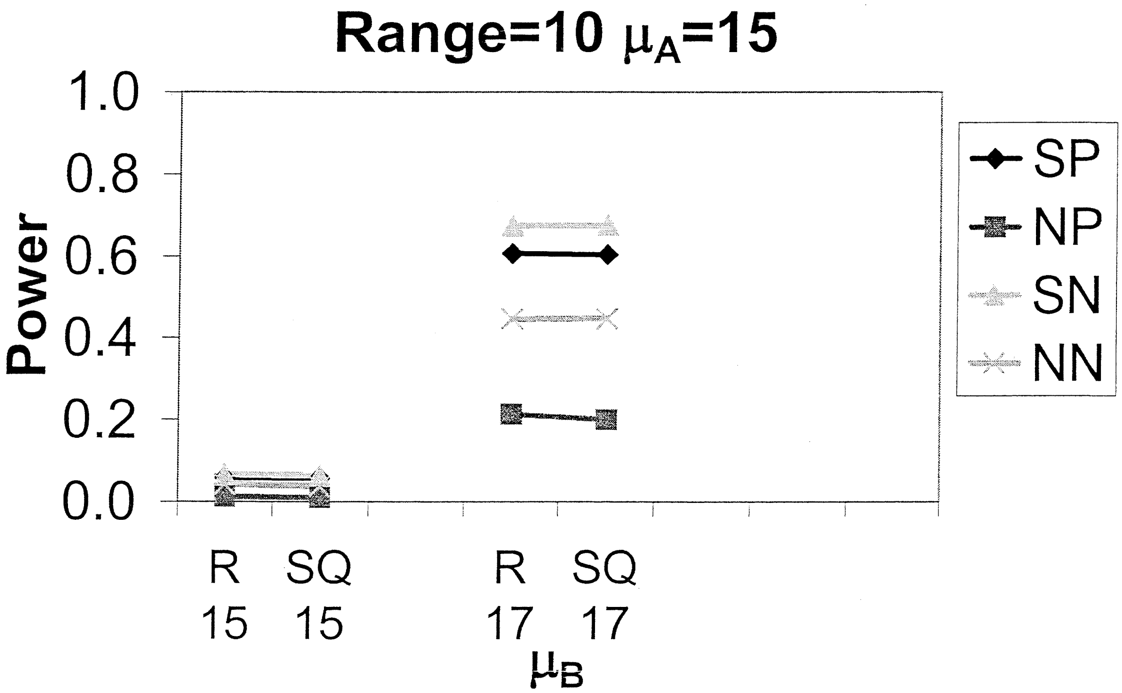


Figure 9. Comparison of power curves of rectangular versus square design for large range and large means.