PAPADAKIS NEAREST NEIGHBOR ANALYSIS OF YIELD IN AGRICULTURAL EXPERIMENTS

Radha G. Mohanty

PAPADAKIS NEAREST NEIGHBOR ANALYSIS OF YIELD IN AGRICULTURAL EXPERIMENTS

Radha G. Mohanty
Monsanto

Abstract: Papadakis analysis, originally proposed by Papadakis in 1937 belongs to a larger class of methodologies called the nearest neighbor analysis which is primarily based on the fact that plots in close proximity ("neighbors") are exposed to similar environmental conditions and therefore, for a given plot, information from its neighboring plots could be used for adjustment of its response for spatial variability. The basic theory behind the application of Papadakis methodology to field trials is relatively simple. It is based on an analysis of covariance where the covariate is an index of fertility (environment), and the response is some observable trait (e.g., grain yield), which is adjusted up or down to reflect the effect due to spatial variability. There have been several references in the literature to application of Papadakis methodology to field trials where the analysis is routinely carried out on data coming from a replicated design within a testing location. The application that is presented here is an exception to the rule in that the analysis is conducted on multi-location data with single replication per location. In plant breeding industry, a recent trend has been to move towards one-replicate testing system to maximize the coverage of the testing environments. Note that for a one-replicate test, no design such as a Lattice, can be used for adjustment of the observations for spatial variability. We start with describing the theory and methodology behind the proposed Papadakis analysis for multi-location data. Several practical problems such as impact of missing values on Papadakis covariate, choice of homogeneous vs. heterogeneous slope coefficient, and effect of influential observations, etc. are discussed and solutions are proposed. Finally, results from several validation studies on corn yield data, including comparison to lattice adjusted plot values and ANOVA on adjusted vs. unadjusted data are presented to demonstrate the benefit from the proposed procedure.

Keywords: *Papadakis analysis, NNA, Nearest neighbor adjustment, Spatial analysis, Genotype by environment interaction, Plant breeding, Lattice analysis.*

1. Introduction

Application of spatial methods to the design and analysis of agricultural experiments is based on the existence of spatial autocorrelation among neighboring plots at a testing location. In the following discussion, we shall use the term "test" and "location" interchangeably to denote a testing location with either one or multiple replications. Traditionally, the designs for plant breeding experiments have been complete or incomplete blocks, where the means for local control is blocking of experimental units. The inherent assumption behind successful

blocking is that the fertility pattern within the test is regular and is known to the experimenter. However, these assumptions rarely hold in practice for reasons such as irregular or patchy fertility pattern, presence of row by column interaction, or extremely large size of the block to accommodate a large number of genotypes tested in early generation screening trials. It is important to realize that in a trial, the more elaborate the chosen method of local control, the worse the consequences are if it fails to achieve its goals (Pearce, 1998).

2. Spatial and Nearest Neighbor Methods

One of the objectives of incorporation of spatial methods at the design or the analysis stage in a plant breeding trial is similar to that of using local control in the traditional design of experiments. However, in many cases, spatial methods of analysis have the advantage of being independent of the design used in the trial (Bartlett, 1978). Nearest Neighbor Analysis (NNA) methods are spatial methods which rely on the fact that experimental plots in close proximity ("neighbors") are exposed to similar environmental conditions, and therefore, for a given plot, information from its neighbors can be used for adjustment of its response to account for the spatial variation. The benefit of using NNA in yield trials is extensively discussed in the literature (Bhatti *et. al.*, 1991; Ball *et. al.*, 1993), and the superiority of NNA over the traditional Randomized Complete Block (RCB) is now well known (Stroup *et. al.*, 1994). In fact, it has recently been shown that spatial methods such as NNA can provide more accurate and precise estimate of genotype effect than either complete or incomplete block analyses (Cullis *et. al.*, 1998; Zimmerman and Harville, 1991; Wu and Dutilleul, 1999).

The mathematical basis for NNA is the simple Markovian model:

$$Y_r = \rho Y_{r-1} + \varepsilon_r, \qquad (0.1)$$

where, Y_r is the response from the r^{th} plot, Y_{r-1} is the response from the neighboring plot (without loss of generality, say, to the left of Y_r), and ε_r is the residual which is uncorrelated with Y_{r-1}. It is clear that the closer ρ is to 1, the better the performance of NNA (Bartlett, 1978).

3. Papadakis Analysis

Papadakis analysis (Papadakis, 1937) is a member of NNA based on the theory of Analysis of Covariance (ANCOVA) where the covariate is an index of fertility (environment) and the response variable is some observable trait (e.g., grain yield) which is adjusted up or down to reflect the effect due to spatial variability. As with any other NNA method, Papadakis analysis has been routinely applied to data within a trial (location) coming from a replicated design (e.g., RCB). Typically, the procedure involves an ANCOVA where the covariate corresponding to an observation is constructed using the mean of the residuals of the neighboring plots. The definition of neighboring plots will be more explicit in later sections.

4. One Replicate Testing System

A recent trend in the agricultural industry has been to switch to testing programs that involve only one replication per location. The intent is to reallocate the plant breeding resources so that the number of testing environments (locations) is maximized. In plant breeding literature (Johnson *et. al.*, 1992), it is known that conducting yield trials at multiple locations is an essential component of any breeding program for developing cultivars with stable performance across a broad range of growing locations. In fact, it can be shown that if the cost of an additional location is not excessively high relative to that of an additional replication, and the error variance is low compared to the variance of genotype by environment interaction (GXE), then one-replicate testing actually results in a lower standard error for the cultivar mean (Dofing and Francis, 1990).

In a testing system with single replication per location, the need for spatial methods is even greater since no traditional complete or incomplete block designs (e.g., lattice designs) can be used for adjustment of cultivar means for spatial variations. For the same reason, Papadakis or any other traditional NNA methods, that require replication, can not be applied to data within a testing location with one replication. Historically, all applications of Papadakis method to yield trials have been within a testing location in the context of replicated designs. In fact, use of any spatial analysis that goes across trials over space and time has not been addressed in the literature (Brownie *et. al.*, 1993). The objective of the present study is to propose a Papadakis analysis method that goes across testing locations with one replication per location.

5. Across-Location Papadakis Analysis

In plant breeding industry, a collection of testing locations, each with the same set of genotypes, is often referred to as an "experiment". The method that we propose here is applied to data coming from multiple locations of an experiment with one replication per location. Therefore, for a given trait, the structure of the input data set for the analysis is a two-way location by genotype layout with one observation per cell. This data structure corresponds to the basic additive model:

$$y_{ij} = \mu + h_i + l_j + \varepsilon_{ij}, \qquad (0.2)$$

where y_{ij} is the observation from the i^{th} genotype at the j^{th} location, μ is the overall mean effect, h_i is the effect of the i^{th} genotype, l_j is the effect of the j^{th} location, and ε_{ij} is the residual. To incorporate Papadakis analysis, a term involving the Papadakis covariate needs to be introduced into model (0.2), following which an ANCOVA can be conducted. Similar to the traditional intra-location Papadakis analysis, the covariate is constructed by taking the mean of the residuals corresponding to the neighboring plots of the observation y_{ij}.

Spatial Trend and GXE:

In model (0.2), the residual ε_{ij} consists of three components: (i) spatial trend, (ii) GXE, and (iii) random error. Ideally, Papadakis covariate should be constructed exclusively from the pure trend part of the residuals to avoid "contamination" of the covariate by GXE. However, separation of the GXE and the spatial trend part is not a trivial problem since effect due to GXE could manifest itself as spatial trend and vise versa. To deal with this problem, in the developmental phase of this study, an iterative method was developed. The method was a combination of NNA and the method of Additive Main Effects and Multiplicative Interaction (AMMI) (Zobel *et. al.* , 1988). This iterative procedure, which we shall refer to as NN-AMMI, consisted of the following steps:

1. Fit an AMMI_t model to raw data: $y_{ij} = \mu + h_i + l_j + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \varepsilon_{ij}$,
   where λ_k is the k^{th} singular value of the matrix of the residuals, and α_{ik} and γ_{jk}, $k = 1,2$ are the k^{th} PCA axis with respect to hybrid and location, respectively.
2. Using the residuals ε_{ij} from Step 1, construct the Papadakis covariates x_{ij}.
3. Define $y_{ij}^{*} = y_{ij} - \sum_{k=1}^{t} \hat{\lambda}_k \hat{\alpha}_{ik} \hat{\gamma}_{jk}$ (i.e., separate the GXE part).
4. Fit the NN model to y_{ij}^{*}, i.e., compute $y_{ij}^{*} = \mu + h_i + l_j + \beta x_{ij} + \varepsilon_{ij}^{*}$.
5. Compute $y_{ij}^{**} \equiv y_{ij} - \hat{\mu} - \hat{h}_i - \hat{l}_j - \hat{\beta} x_{ij}$ ($\hat{\mu}, \hat{h}_i, \hat{l}_j, \hat{\beta}$ are from Step 4).
6. Fit the multiplicative model: $y_{ij}^{**} = \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \varepsilon_{ij}^{**}$.
7. Define $u_{ij} = y_{ij} - \hat{\mu} - \hat{h}_i - \hat{l}_j - \sum_{k=1}^{t} \hat{\lambda}_k \hat{\alpha}_{ik} \hat{\gamma}_{jk}$ using $\hat{\mu}, \hat{h}_i, \hat{l}_j$ from Step 4 and $\hat{\lambda}_k, \hat{\alpha}_{ik}, \hat{\gamma}_{jk}$ from Step 6.
8. Using u_{ij}, construct a new set of values for the covariate x_{ij}.
9. Loop back to Step 3 and Step 4 using parameter estimates from Step 6 and x_{ij} from Step 8.
10. Continue looping until estimates converge.

Theoretically, it seemed logical that the above iterative procedure would gradually refine the Papadakis covariate to its "purest" form freeing it from the effects due to the multiplicative factor, i.e., the GXE interaction. However, after applying this method to corn yield data from Monsanto's testing system, it was observed that even if the procedure converged after only two or three iterations, the value of the slope coefficient β hardly changed after the first iteration. One possible explanation for this phenomenon is that AMMI was unable to distinguish between the trend and the GXE part and was therefore erroneously including the trend component into modeling of the multiplicative interaction. This real data

analysis suggested that NN-AMMI was no better than the pure NNA method. The conclusion then was to revert back to the original non-iterative Papadakis method.

Homogeneous Slope Model:

The basic homogeneous slope form of the across-location Papadakis method, which we propose, is given by the following model:

$$y_{ij} = \mu + h_i + l_j + \beta \cdot x_{ij} + \varepsilon_{ij}, \qquad (0.3)$$

where x_{ij} is the covariate corresponding to the observation y_{ij}, β is the slope coefficient, and the remaining terms μ, h_i, l_j, and ε_{ij} have been defined before in Equation (0.2). Similar to the intra-test Papadakis method, an ANCOVA can be performed using the model (0.3) for obtaining adjusted genotype means and for testing contrasts on genotypes. Alternatively, the approach that we follow in this discussion is consistent with the industry practice where genotype by location values are stored in a database for further analyses to make inferences on varietal selection.

Following (0.3), the adjustment of genotype values are done by the formula:

$$y_{ij}^{adj} = y_{ij} - \hat{\beta} \cdot x_{ij} \qquad (0.4)$$

where y_{ij}^{adj} is the adjusted value and $\hat{\beta}$ is the estimate of the slope coefficient.

6. Model Related Issues

There are at least two major model related issues which need to be resolved for building model (0.3). They are (a) appropriate construction of Papadakis covariate, and (b) possible heterogeneity of β with respect to location and/or genotype.

(a) Construction of the Papadakis covariate: There are two choices for constructing a Papadakis covariate, which appear the most in the literature. They are (i) two-neighbor covariate and, (ii) four-neighbor covariate. The four-neighbor covariate is based on the residuals from all four neighbors (east, west, north, and south directions), where as the two-neighbor covariate uses residuals either from the east and west neighbors or from the north and south neighbors. During the course of this study, several validation studies using corn yield data from Monsanto's testing system were conducted on all of the above alternative choices of covariate construction. The four-neighbor covariate produced the most precise spatial adjustment of the observations (see the discussion in the Validation Studies section).

(b) Heterogeneity of slope coefficient: The assumption of homogeneous slopes is routinely tested in general ANCOVA but does not appear to be tested in application of Papadakis analysis found in the literature. One explanation may be that, traditionally Papadakis analysis has been exclusively applied to within-location RCB data where the number of replications is usually small. Therefore, a test of heterogeneity with respect to treatment would have very few data points for each level of the treatment and consequently, very low power. As for replication

by covariate interaction, typically in an RCB, it is assumed that replication by treatment interaction is non-significant, so this might suggest an assumption of no replication by covariate interaction. Further, if the assumption of homogeneity with respect to treatment is not routinely tested in Papadakis, it probably follows that the assumption of homogeneity with respect to block would not be tested either, as the former is usually routine in ANCOVA whereas the latter is not.

In our present application though, the situation is much different. First, there usually are a large number of "blocks" (approximately 20-35 locations in an experiment), so testing the assumption of homogeneity with respect to the treatment (genotype) would probably not lack power. Second, it is usually assumed that our blocks do interact with our treatment (GXE), so it might be reasonable to assume heterogeneity of the slope with respect to our blocking factor (location) also.

As noted previously, we are interested in the comparisons among genotypes and not among the locations. If a data set is heterogeneous with respect to location and homogeneous with respect to hybrid, we can adjust the observation within each location separately using different slope coefficients for different locations and still compare the genotypes. Furthermore, it can be shown that we can compare the genotypes for any subset of the locations (Hendrix *et. al.*, 1982).

Unlike location, heterogeneity with respect to genotype introduces some problems into the implementation of the method. In that case, the value of the covariate at which two genotypes are compared can alter the resulting conclusion about their relative performance. This would make our proposed use of adjusted data nearly impossible. In fact, the interaction of a covariate and the treatment factor of interest introduces problems in the interpretation similar to those present when two treatment factors interact and conclusions about main effects need to be qualified (Hendrix *et. al.*, 1982).

Using corn yield data on several experiments from Monsanto's testing system, tests of heterogeneity of Papadakis slope coefficient were conducted. It was found that the locations were heterogeneous with respect to the covariate in every experiment. The p-value for the test of homogeneity was almost always equal to zero. On the other hand, test for homogeneity with respect to hybrid was significant in approximately 25% of the experiments using a significance level of .05. However, the size of the F-statistic for heterogeneity with respect to hybrid was usually between 1 and 2, whereas the F-statistic for location heterogeneity was much larger. Given that heterogeneity with respect to location is more prevalent and does not cause the same problems in interpretation that heterogeneity with respect to hybrid does, from point of view of practical application, we propose a model that routinely assumes heterogeneity with respect to location and homogeneity with respect to genotype.

Final Model:

The final model for across-location Papadakis analysis now becomes:

$$y_{ij} = \mu + h_i + l_j + \beta_j \cdot x_{ij} + \varepsilon_{ij,} \qquad (0.5)$$

where β_j is the slope for the j^{th} location, and the remaining terms in the model are as defined before in (0.3). As in (0.4), the adjusted response using this model is given by

$$y_{ij}^{adj} = y_{ij} - \beta_j^{est} \cdot x_{ij}, \qquad (0.6)$$

where β_j^{est} is the estimate of the slope coefficient for the j^{th} location.

7. Implementation Issues

In the following paragraphs, some of the most important implementation issues will be discussed.

***(a) Missing values and the criterion of adequate neighbors**:* Missing values for the Papadakis analysis are troublesome in that not only is the plot, whose yield is unobserved, affected but so also are the neighboring plots whose covariate cannot be constructed using all four neighbors.

Examination of the single-replicate yield trial data from the historical Monsanto testing system suggested that approximately 30% of the tests each year had at least one missing plot. Among the tests with at least one missing plot, most had five or fewer missing plots. One should be concerned with not just the number of missing plots but also the pattern in which they are missing or the ultimate effect on the analysis. In other words, missing values could be positioned throughout a test in such a way that every interior plot still has at least 3 neighbors, or they could be positioned such that many plots has two or fewer neighbors. Note that, for a given plot, having 3 or 4 neighbors is preferable to having 1 or 2 neighbors because then the information about spatial variability will assuredly be available in both vertical and horizontal directions. Also note that the border plots have at most 3 neighbors and the corner plots have at most 2 neighbors.

Lack of replication means that a design based missing value estimation procedure can't be implemented at the location resolution. A practical solution to this problem is to determine some objective criterion in terms of data adequacy and then drop locations from the analysis which do not meet that criterion. In the following paragraph, we present an example where such a criterion is established in the case of an experiment containing tests of size 7x7, i.e., tests with 49 plots.

Application to 7X7 test:

To determine a rule for the inclusion or exclusion of a 7x7 test with respect to having at least three neighbors for a sufficient number of its interior (non-border) plots, several real data analyses were conducted using historical data from Monsanto's corn testing system. The first analysis showed that, for tests with 4 missing non-corner plots, the average number of neighbors per plot was 3.2 and for the tests with 5 missing non-corner plots, the average number of neighbors per plot was less than 3. As mentioned before, the presence of a minimum of three neighbors for a plot guarantees that its covariate will use spatial information in both the vertical and horizontal directions. In a second analysis, only the tests that had at least one non-corner plot with less than 3 neighbors were considered. Then, for such a test, the average number of plots with less than 3 neighbors for different numbers of missing non-corner plots was considered. It was found that, for the case of tests with 4 missing non-corner plots, the average number of plots with fewer than 3 neighbors was 5.14. The results from the above two

analyses jointly point in the direction of not using a test which has 5 or more non-corner plots, each with less than 3 neighbors. It is worth mentioning here that since the corner plots can have at most 2 neighbors, a test with 5 non-corner plots, each with less than 3 neighbors, will have a total of 9 plots with fewer than 3 neighbors.

(b) Identification of influential observations: Use of the model (0.5) with slope heterogeneous with respect to location implies that the number of data points that essentially determines the slope coefficient is equal to the number of genotypes at a location. This means that a single data point could be quite influential in the determination of the slope coefficient for a given location. Therefore, it seems reasonable to screen for highly influential data points and remove them prior to the estimation of the slope parameter. Note that the observations from those deleted plots can be ultimately adjusted using the fitted model based on the data from the remaining plots.

In practice, for identification of influential observations, the diagnostic statistic, DFFITS (SAS PROC REG) is recommended since it highlights points influential in the estimation of both the slope and the intercept parameters. Before construction of the across-location model (0.5), a regression line is fit to data from each individual location, regressing yield (adjusted for the genotype effect) on the Papadakis covariate and then, all data points with a DFFITS value greater than some predetermined threshold are dropped before the model fitting is done.

(c) Negative Papadakis slope: Occurrence of negative slope in Papadakis analysis is not very intuitive since it implies, for example, that the response from a plot with a large covariate (i.e., in a higher yielding region) will be adjusted up and the response from a plot with a small covariate (i.e., in a lower yielding region) will be adjusted down. As a possible explanation of this phenomenon, we suggest inter-plot competition (Kempton and Howes, 1981). Note that, in the presence of inter-plot competition, a low-yielding plot will be associated with higher yielding neighboring plots, and a high-yielding plot will be associated with a lower yielding neighboring plots. We claim that negative slopes indicate smaller spatial trend relative to inter-plot competition. To justify the claim, note first of all that at all locations, there exists some degree of spatial trend and some degree of inter-plot competition. Whichever of these two components of the residual produces a stronger 'signal' determines the sign of the Papadakis slope. It is reasonable to assume that for a given experiment, the amount of inter-plot competition is approximately constant across all tests since they all contain the same set of genotypes. Conversely, within-test trend or the spatial variability varies from test to test. The net effect of the approximately constant background inter-plot competition and the varying trend component is reflected in the Papadakis slope coefficient; the slope is large and positive when trend is significantly stronger than inter-plot competition and decreases proportionally as the trend gets weaker. In the extreme, little or no spatial trend leads to Papadakis slopes which have large negative values reflecting presence of only inter-plot competition.

The hypothesis that inter-plot competition exists to some degree in all tests is further substantiated by the following analysis. Using raw data from 55 experiments from Monsanto's testing system, Papadakis slopes were computed using covariates based only on row (east-west) neighbors and then using covariates based only on column (north-south) neighbors. In 54 out

of 55 experiments, the mean slopes (across tests) using the row (east-west) neighbors were smaller than the mean slopes using the column (north-south) neighbors. Note that, given the extremely elongated shape of the testing plots, the distance between adjacent plots in the east-west direction is about one-fourth of the distance in the north-south direction and therefore, the inter-plot competition is expected to be much stronger between the row neighbors than that between the column neighbors. The above relationship between the two types of slopes held up both in the tests with positive slopes and tests with negative slopes. In other words, a similar degree of inter-plot competition was present in all tests, both with positive and negative Papadakis slopes, suggesting that only the amount of trend varied from location to location.

Another piece of evidence suggesting the presence of inter-plot competition and its direct relationship to Papadakis slope was found by analyzing several years of data from all experiments with 2-replicate 7X7 lattice tests in Monsanto's testing system. The tests were first analyzed using lattice analysis on the 2-replicate data and then using the proposed across-location Papadakis analysis with data from one replication at a time. After completion of the analyses, tests were divided into two groups, one with positive Papadakis slope and the other with negative Papadakis slope, and then the lattice relative efficiency (R.E.) for both types of tests were examined. In theory, in the presence of inter-plot competition, the lattice analysis should not perform as well in tests with less trend as those with more trend. That is, the lattice block adjustments can more effectively adjust out trend, particularly if it occurs along blocks, than the inter-plot competition which most likely does not occur along blocks but rather at a higher resolution. Therefore, it could be hypothesized that the R.E. of lattice would be less in tests with negative Papadakis slopes than those with positive Papadakis slopes. The data indeed supported this hypothesis. It was observed that the average R.E. for tests with positive slopes was 127 whereas that for tests with negative slopes was 114.

The results of these analyses seem to suggest that inter-plot competition always reduces the Papadakis slope coefficient and sometimes to such an extent that the Papadakis slope becomes negative.

8. Validation Studies

Several validation studies to assess the validity of the proposed method were undertaken using historical and current yield data from Monsanto Corn Research. Some of the studies required data from multi-replicate testing system, which were available in Monsanto's corn testing system from past several years. Details of three of the validation studies are discussed below.

(a) Comparison with lattice-adjusted plot values: It appears that one of the natural ways of assessing the proposed Papadakis analysis would be to consider data from 2-replicate lattice designs and then compare the lattice adjusted plot values on one of the replications to the Papadakis adjusted values on that replication. Note that lattice analysis uses data from both replications for computing lattice adjusted plot values and the lattice adjusted mean of a genotype is equal to the mean of the two lattice adjusted plot values corresponding to the two

replications for that genotype. On the other hand, the proposed Papadakis analysis is based on single replicate data only. Therefore, if we find that the adjustment factors (adjusted value minus raw data) from Papadakis analysis are similar to those from lattice analysis, it would certainly be an evidence in favor of Papadakis analysis since it requires only half the amount of data compared to the lattice analysis. The results of the data analysis indeed supported the hypothesis. The average correlation between the adjustment factors from Papadakis analysis and lattice analysis was 0.74.

(b) Cross-Validation: For a given hybrid from a 2-replicate test, we can think of the two observed values coming from the two replications as two different estimates of the true yield of the hybrid. Therefore, it appears logical to use one of the replications as a predictor of the other replication. Without loss of generality, we can use the second replication to predict the first replication. For using the second replication as a predictor, we can either use the observation (raw data) from the second replication itself or we can do Papadakis adjustment only on the second replicate and then use the adjusted second replicate data as the predictor of the first replication. The claim is that, between the two predictors, the raw rep 2 and the adjusted rep 2, the adjusted rep 2 value will be a "better" predictor of the raw rep 1 value. The word "better" here is used in terms of the magnitude of mean squared prediction error (MSPE).

From Monsanto's corn research, data from all experiments from years 1993 through 1995 were used for this study. For a given year, experiments with 2-replicate lattice tests were taken and Papadakis adjustment was done on the second replication. Recall that Papadakis analysis is done by experiment. As explained in the previous paragraph, MSPE was computed at the test level first using raw rep 2 as a predictor of raw rep 1 ("raw method"), and second, using adjusted rep 2 as a predictor of raw rep 1 ("adjusted method"). Then at the test level, the percent reduction in MSPE from the raw method to the adjusted method was calculated. The percent reduction in MSPE was then averaged first across tests up to the experiment level and then averaged across experiments to get an overall figure for the corresponding year.

The following table summarizes the results. The average percent reduction across the three years is weighted by the number of experiments in a year.

Percent Reduction in MSPE from Raw to Adjusted Method

Year	Number of Experiments	% Reduction in MSPE
1993	58	7.74
1994	18	5.73
1995	8	6.12
Wt. Avg.		7.16

(c) ANOVA: This analysis required only one-replicate data where we used the additive model with hybrid and location, first on the raw data and then on the Papadakis adjusted data. The MSE's coming from the two ANOVA's were then compared. This method has an added appeal

because it is identical to the analysis involved in the current industry practice of "experiment processing", which is an ANOVA on location by hybrid values used for selection of hybrids.

For this study, all experiments with one-replicate tests from years 1996 through 1998 were taken from Monsanto's corn research. For a given experiment in a year, an ANOVA was conducted based on the additive model (0.2) of test and hybrid using two alternative data sets: (a) raw data and, (b) Papadakis adjusted data. Using the MSE's from the two alternative ANOVA's, reduction in MSE from the ANOVA with raw data to the ANOVA with adjusted data was calculated. These percent reductions were then averaged across experiments for a year.

The following table summarizes the results. The average across years is weighted by the number of experiments in the corresponding year.

Percent Reduction in MSE from the ANOVA with Raw Data to the ANOVA with Adjusted Data

Year	Number of Experiments	%Reduction in MSE
1996	55	12.6
1997	56	11.3
1998	62	14.3
Wt. Avg.		12.8

The three validation studies discussed above certainly add credence to the usefulness of the proposed procedure as a means for adjusting the data for spatial variation from one replicate testing system. Figures 1 and 2 provide contour plots of corn yield residuals (using model (0.2)) before and after Papadakis adjustment in a 7x7 test from Monsanto testing system. Note that the spatial pattern in the residuals using raw data is not evident in the residuals after Papadakis adjustment is done.

9. Conclusion

In recent years, the trend in the plant breeding industry has been to move towards testing systems having one replication per location with an objective of reallocating the breeding resources to maximize the number of testing environments. In this study, we have shown that in experiments containing one-replicate tests, the benefits from Nearest Neighbor Papadakis analysis can still be achieved if the traditional intra-location analysis is extended to the proposed across-location procedure. Between the two choices of (a) analyzing raw data (current industry practice) and (b) analyzing Papadakis adjusted data, the latter is shown to provide more precision in comparison of genotype means. The real data validation analyses discussed in this study have been done on corn yield only. However, it is expected that the proposed method would produce similar benefits for other crops and any other trait as long as

the trait under consideration has the tendency to respond explicitly to variation in spatial characteristics of the field.

References

Ball, T. B., Mulla, D. J., and Konzak, C. F. (1993), "Spatial heterogeneity affects variety trial interpretation," *Crop Science*, 33, 931-935.

Bartlett, M. S. (1978), "Nearest neighbor models in the analysis of field experiments," *J. R. Statist. Soc.*, B, 40, 2, 147-174.

Bhatti, A. U., Mulla, D. J., Koehler, F.E., and Gurmani, A. H. (1991), "Identifying and removing spatial correlation from yield experiments," *Soil Sci. Soc.Am. J.,* 55,1523-1528.

Brownie, C., Bowman, D. T., and Burton, J.W. (1993), "Estimating spatial variation in analysis of data from yield trials: a comparison of methods," *Agron. J.*, 85, 1244-1253.

Cullis, B., Gogel, B., Verbyla, A., and Thompson, R. (1998). "Spatial analysis of multi-environment early generation variety trials", *Biometrics*, 54, 1-18.

Dofing, S. M. and Francis, C. A. (1990), "Efficiency of one-replicate testing," *J. Prod. Agric.*, 3, 399-402.

Hendrix J. H., Carter M. W., Scott D. T. (1982). "Covariance analysis with heterogeneity of slopes in fixed models", *Biometrics*, 38, 641-650.

Johnson, J. J., Alldredge, J. R., Ullirich, S. E., and Dangi, O. (1992). "Replacement of replications with additional locations for grain sorghum cultivar evaluation", *Crop Science*, 32, 43-46.

Kempton, R. A. and Howes, C. W. (1981), "The Use of Neighboring Plot Values in the Analysis of Variety Trials", *Applied Statistics*, Vol. 30, 1, 59-70.

Papadakis, J. S. (1937), "Methode statistique pour des experiences sur champ.," *Bull. Inst. Amel. Plantes a Salonique*, No. 23.

Pearce, S. C. (1998), "Field experimentation on rough land: the method of Papadakis reconsidered," *Journal of Agricultural Science*, 131, 1-11.

Stroup, W.W., Baenziger, P. S., and Mulitze, D.K. (1994), "Removing spatial variation from wheat yield trials: a comparison of methods," *Crop Science*, 86, 62-66.

Wu, T. and Dutilleul, P. (1999), "Validity and efficiency of neighbor analyses in comparison with classical complete and incomplete block analyses of field experiments," *Agron. J.*, 91, 721-731.

Zimmerman, D. L. and Harville, D. A. (1991), "A random field approach to the analysis of field-plot experiments and other spatial experiments," *Biometrics*, 47, 223-239.

Zobel, R.W., Wright M.J., and Gauch, Jr., H.G. (1988), "Statistical analysis of a yield trial," *Agronomy Journal*, 80, 388-393.

Figure 1: Residuals before Papadakis adjustment

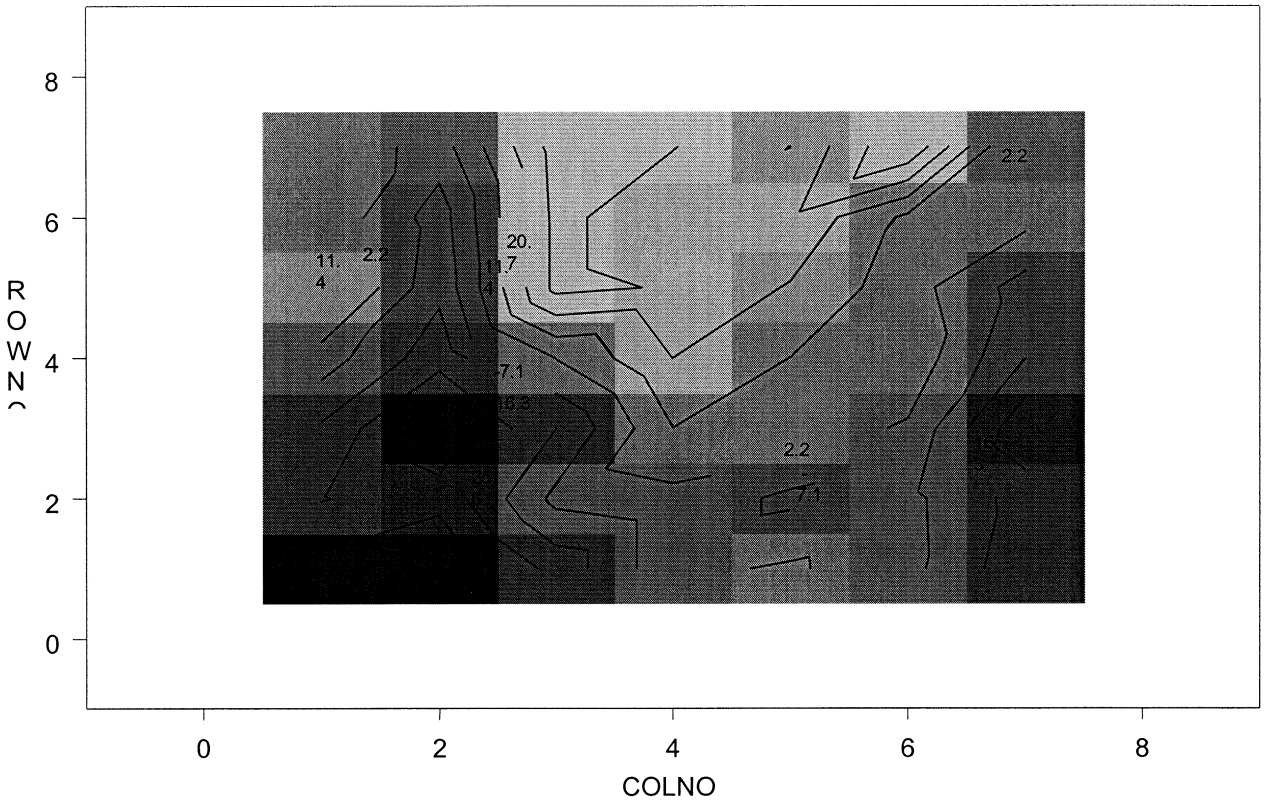

Figure 2: Residuals after Papadakis adjustment