ABSTRACT

"A Sample Homogeneity Test for Saskatchewan Grain Farm Data: An Application of the Principal Components Technique"

Malcolm B. Cairns and Ihn H. Uhm*

This paper demonstrates the use of principal component analysis to systematically identify aberrant observations in a multi-dimensional context to ensure sample homogeneity. Prior to quantitative analysis based on survey data such as diversified Saskatchewan grain farms, a careful examination of the data using this technique is illustrated.

*Railway Transport Committee and Research Branch respectively, Canadian Transport Commission, Ottawa, Canada.

**Selected paper prepared for presentation at the session on "Farm Size and Structure", of the American Agricultural Economics Association Annual Meeting, August 4, 1982, Utah State University, Logan, Utah.

A Sample Homogeneity Test for Saskatchewan Grain Farm Data: An Application of the Principal Components Technique

by

Malcolm B. Cairns
Railway Transport Committee
Canadian Transport Commission
Ottawa, Ontario

and

Ihn H. Uhm
Research Branch
Canadian Transport Commission
Ottawa, Ontario

*The views and opinions expressed herein are solely those of the authors and do not necessarily reflect those of the Canadian Transport Commission or the Government of Canada. Mr. Andrew W. Gemmell reviewed an earlier draft and offered helpful comments which improved the paper.

A Sample Homogeneity Test for Saskatchewan Grain Farm Data: An Application of the Principal Components Technique

Malcolm B. Cairns and Ihn H. Uhm*

Introduction

The availability of more and better micro-economic data for farms have increased in recent years and they are now readily available through Canfarm Service Agency Ltd., Guelph, Ontario. Research opportunities into micro-economic issues, therefore, have increased correspondingly. Technical efficiency, for example, has been a continuous research interest among agricultural economists in Canada but until now, empirical studies have been rather limited due to data constraints. However, using farm level survey data, the researcher must be concerned about sample homogeneity. This is especially true when the industry to be analyzed is fragmented by its very nature or individual decision units (farms) are distinctly different from each other. Such variability can arise due to physical, economical and managerial constraints. This is the case with Saskatchewan grain farms, the subject studied in this paper. In this case, a sample sufficiently homogeneous to do meaningful economic analysis, was difficult to obtain. It is the authors' contention that a systematic test procedure, to ensure a representative sample such as principal components analysis is an appropriate tool for examining an unstructured sample of data.

The objective of this paper, therefore, is 1) to demonstrate the principal components technique as a device to identify aberrant observations in the farm level data and 2) to discuss the consequences of including or not including such multivariate outliers in the sample.

*Canadian Transport Commission, Ottawa, Canada.

Characteristics of Saskatchewan Grain Farms

The grain production industry in Saskatchewan is composed of a large number of individual farms located on either brown, dark brown or black soils (see Figure 1). There were about 70 thousand such farms in 1977 and 69 thousand in 1980. Of the 140 million acres of land in Saskatchewan, about 42.1 million acres of cropland were available in 1977 and 24.3 million acres of land were seeded to grains (wheat, oats, barley, rye and oilseeds). The seeded acreage represents roughly 57 per cent of total cropland in Saskatchewan while the remainder was summer fallow that year. About 73 per cent of the seeded cropland (i.e., 17.7 million acres) was allocated to wheat and the rest was allocated to other grains.

(Insert Figure 1)

In 1977, over 41 thousand farms (i.e., 59% of farms) raised cattle, 14 thousand raised pigs and less than one thousand raised sheep. An estimated 61 thousand farms (i.e., 87% of farms) produced wheat with an average yield of 785 kg. (or 28.8 bu.) per acre.

Thus, it can be seen that grain farms in Saskatchewan are, in general, diversified as to their mixture of grains and livestock. In addition, the majority of farms cultivate more than one grain type in a given year even though it was seen that wheat is by far the most popular.

Individual farms have different characteristics such as farm size, crop mix, soil fertility, weather conditions, grain hauling distance to the country elevator, capital intensity, degree of concentation on grain as opposed to livestock, and so forth. These difference characteristics result in inter-farm variation in the selection of cropping programs, in the cost

of production and transportation, in yields realized and in net farm income generated.

For an industry such as grain farms in Saskatchewan, it is a legitimate question as to whether one can do a quantitative economic analysis, based on farm level data, to answer such questions as whether economies of farm size exist. This paper shows a systematic way to screen the data to ensure that the sample is sufficiently homogeneous to carry out such an analysis?

The Data and Methodology

The Data

The data for this paper was obtained from Canfarm Service Agency's, Guelph, Ontario accounting record data base. This was developed from a special survey of 720 farms in Saskatchewan, including 670 grain farms. The geographical distribution of sample farms are shown in Figure 1. The Canfarm records include statements of farm income, cash flow, as well as assets, liabilities and owner's equity. The data was processed in the following steps: (1) the sample farms were grouped by geographical location with respect to brown, dark brown, and black soil zones; (2) seeded and quota acreage by type of crop were added to the data base; (3) the cost components were split into production and transportation.

Returns to family labour, management and equity capital were not included in the Canfarm cost components and no allowance has been made for them. Canfarm's definition of a grain farm was accepted -- a farm which generated more than 50% of its total income from grain. In fact, on average, about 75% of the income was generated from grain production.

Methodology

To investigate empirically whether there was any natural partitioning in the data indicating the sample was not homogeneous, the principal component analysis (Gnanadesikan) technique was adopted as a screening device and was used to examine selected characteristics of the sample farms. Eleven variables were selected which included: size of farm by output; total seeded acreage; total cost; total income; net income; fertilizer application rate; degree of grain specialization; degree of reliance on hired labour; average cost of production per unit of output; degree of specialization on high yielding crops; and equity cost of capital.

Principal component analysis is a general statistical technique which has proven useful in analyzing collections of units -- in this case, Saskatchewan grain farms -- for which a sequence of measurements or characteristics have been recorded with the intent of exposing outlying units and significant clusters of units (Gnanadesikan). The basic idea of principal component analysis is to describe the dispersion of the array of n points in p-dimensional space by introducing a new set of orthogonal linear co-ordinates so that the sample variances of the given points with respect to these derived co-ordinates are in decreasing order of magnitude. Thus, the first principal component is such that the projections of the given points onto it have maximum variance among all possible linear co-ordinates; the second principal component has maximum variance subject to being orthogonal to the first; and so on.

Let us represent the p operating characteristics or variables under investigation as vector y

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_p \end{bmatrix}$$

and then construct a (p x n) matrix Y

$$Y = (\underset{\sim}{y}_1, \underset{\sim}{y}_2, \ldots, \underset{\sim}{y}_n)$$

where n = 670 farms and p = 11 characteristics of each farm.

The definition of the chosen 11 variables are:

y_1 = the size of the i^{th} farm in the j^{th} soil zone by unit of output in kilograms of grain (S_{ij})*;

y_2 = fertilizer application rate of the i^{th} farm in the j^{th} soil zone measured by fertilizer expenses in proportion to the total variable cost in dollars (FE_{ij});

y_3 = reliance on hired labour of the i^{th} farm in the j^{th} soil zone measured by expenses related to hired labour, specialized labour, and custom work in proportion to the total variable cost in dollars ($SLCW_{ij}$);

y_4 = degree of specialization in grain production of the i^{th} farm in the j^{th} soil zone measured by the percentage of total income from grain crops as opposed to livestock (SP_{ij});

y_5 = the percentage of total seeded acres seeded to the high yielding crops, i.e., oats and barley;

y_6 = average production cost per kilogram of output of the i^{th} farm in the j^{th} soil zone in dollars (APC_{ij});

y_7 = total cost in dollars;

y_8 = total income in dollars;

y_9 = net income in dollars;

y_{10} = total seeded acres;

y_{11} = equity cost of capital, at cost, in dollars

*Variable names in brackets refer to definitions of variables used in the later section discussing Economic Implications.

Since the principal components are not invariant under separate scaling of the original co-ordinates, we first scale each row of matrix Y to have unit variance; this is particularly important since y_1 and y_{10} are in units of kilograms and acres, y_7, y_8, y_9, and y_{11} are in dollars, while the other variables are proportions. Next the correlation matrix R is constructed:

$$R = (r_{ij})$$

where

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii} s_{jj}}}$$

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (y_{ik} - \bar{y}_i)(y_{jk} - \bar{y}_j)$$

$$\bar{y}_i = \frac{1}{n} \sum_{k=1}^{n} y_{ik}$$

We now consider the spectral decomposition of the matrix R

$$R = A \Gamma A'$$

where Γ is a diagonal matrix $\Gamma = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ with λ_i the eigenvalues of R such that $\lambda_1 \geq \lambda_2 \geq \ldots > \lambda_p > 0$ and A is an orthogonal matrix $A = (\underset{\sim}{a}_1, \underset{\sim}{a}_2, \ldots, \underset{\sim}{a}_p)$ with $\underset{\sim}{a}_i$ the corresponding eigenvectors of R. The principal components transformation of the data is then given by:

$$Z = A' (Y - \bar{Y})$$

where $\bar{Y} = (\underset{\sim}{\bar{y}}\ \underset{\sim}{\bar{y}}\ \ldots\ \underset{\sim}{\bar{y}})$ and a plot of the first two principal components is given by a plot of the first two rows of Z. Specifically:

$$Z_1 = a_{11} y_1^* + a_{12} y_2^* + \ldots + a_{1p} y_p^*$$
$$Z_2 = a_{21} y_1^* + a_{22} y_2^* + \ldots + a_{2p} y_p^*$$

where y_i^* are the original variables y rescaled to unit variance and shifted to zero mean. The coefficients $\{a_{1i}\}$ and $\{a_{2i}\}$ are useful in interpreting the relative importance of the original variables to the first two principal components while the eigenvalues λ_1 and λ_2 are useful in interpreting the overall importance of the first two principal components themselves: since the sum of all p eigenvalues is p itself the ratio $(\lambda_1 + \lambda_2)/p$ represents the percentage of variation depicted by the first two principal components.

Principal Components Test Results

The coefficients of the first two principal components for the analysis of the 670 farms are given in Table 1 and the corresponding plot of the first two principal components appear in Figure 2. A two-dimensional plot of the first two principal components is useful in highlighting outlying points and any significant clustering of the points. Examination of Figure 2 reveals a number of distinctive or outlying farms (outliers) which have been highlighted in the figure. Otherwise no particular clustering is apparent. Examination of the coefficients in Table 1 indicates that small values of the first principal component, which is a feature of eight of the nine outliers, may be characterized by large values of farm output, total income and total cost. The remaining outlier, with a large value of both first and second principal components, may be characterized by small values of farm output, total income and total cost together with a large proportion of oats and barley and a low proportion of income from crops. Overall, 46% of the variation is depicted by the first two principal components.

(Insert Table 1 and Figure 2)

Based on the analysis, it is to be suggested that i) the eight outliers representing large farms and the one small farm should be removed prior to an

analysis of the cost structure since they represent extremes of farm size and may distort the analysis of average cost; ii) no clustering of the remaining farms is apparent and since the percentage of variation depicted by the first two principal components is not very high it suggests that the farms are scattered fairly uniformly throughout the range of the variables under consideration.

Economic Implications

To demonstrate the significance of 9 outliers and their influence, an economic model was chosen to compare the estimated parameters by including and excluding 9 outliers. Fleming and Uhm's model on average cost of production for Saskatchewan is chosen for the purpose and the specification of the model is:

$$APC_{ij} = \beta_0 + \beta_1 S_{ij} + \beta_2 FE_{ij} + \beta_3 SLCW_{ij} + \beta_4 SP_{ij} + \beta_5 CEI_{ij} + e_{ij}$$

where CEI_{ij} = cost effective index of the i^{th} farm in the j^{th} soil zone which is defined as the predicted average cost per seeded acre divided by the actual average cost per seeded acre;

e_{ij} = random error term;

APC_{ij}, S_{ij}, FE_{ij}, $SLCW_{ij}$, and SP_{ij} are defined earlier; and β_i are coefficients to be estimated.

Fleming and Uhm hypothesized that the variation of average production cost in Saskatchewan would be inversely related to farm size (S_{ij}), positively related to the fertilizer application rate (FE_{ij}), to the degree of reliance on hired labour ($SLCW_{ij}$), and negatively related to the extent of grain specialization (SP_{ij}), and the level of management efficiency (CEI_{ij}).

The impact of including or excluding the nine identified outliers is shown in Table 2 wherein the data used to estimate Equation 1 includes the data from these nine outlying farms and Equation 2, using the identical model, excludes the outliers. It is to be noted that in comparing two equations, the coefficients for farm size (S_{ij}) and the cost effectiveness index (CEI_{ij}) are substantially larger in Equation 1.

To illustrate visually the consequences of including the nine outliers in the model, a two-dimensional diagram is drawn by eliminating all independent variables except farm size. As shown in Figure 3, the slopes of the two average production cost curves are quite different. The curve derived from Equation 1 shows economies of size even beyond the size of 600,000 kg. of output.

(Insert Table 2 and Figure 3)

The implications of acting on one equation versus the other are immediately different. If the equation including the entire sample (i.e., with the nine outliers - one small and eight very large farms) is used, there is a tendency to understate the average production costs for a significant portion of the population and perhaps therefore to implement policies not in keeping with the population's needs. The technique has shown that it would be better to subdivide the population, and recognize the existence of very large farms with apparently significantly lower average production costs, which may need a separate policy from the remainder of the population.

Conclusions

The principal components technique is a useful tool to systematically detect aberrant observations in the framework of a multi-dimensional pers-

pective. As more microeconomic data becomes available, more studies of the economic behaviour of firms are feasible. However, as is shown here, prior to quantitative analysis based on survey data -- such as diversified Saskatchewan grain farms -- a careful examination of the data using principal components analysis is a most appropriate step.

References

Fleming, Marion S., and Ihn H. Uhm, Impact of Rail Rationalization Proposals on the Net Income Position of the Grain Producer in Saskatchewan, Research Report No. 10-81-02, Research Branch, Canadian Transport Commission, Ottawa, August 1981.

Gnanadesikan, R., Methods for Statistical Data Analysis of Multivariate Observations, New York, John Wiley and Sons, 1977.

Saskatchewan Agriculture, Farm Business Review for the Year 1977, Regina, Saskatchewan, 1978.

Saskatchewan Agriculture, Agricultural Statistics 1980, Regina, Saskatchewan, 1980.

FIGURE 1

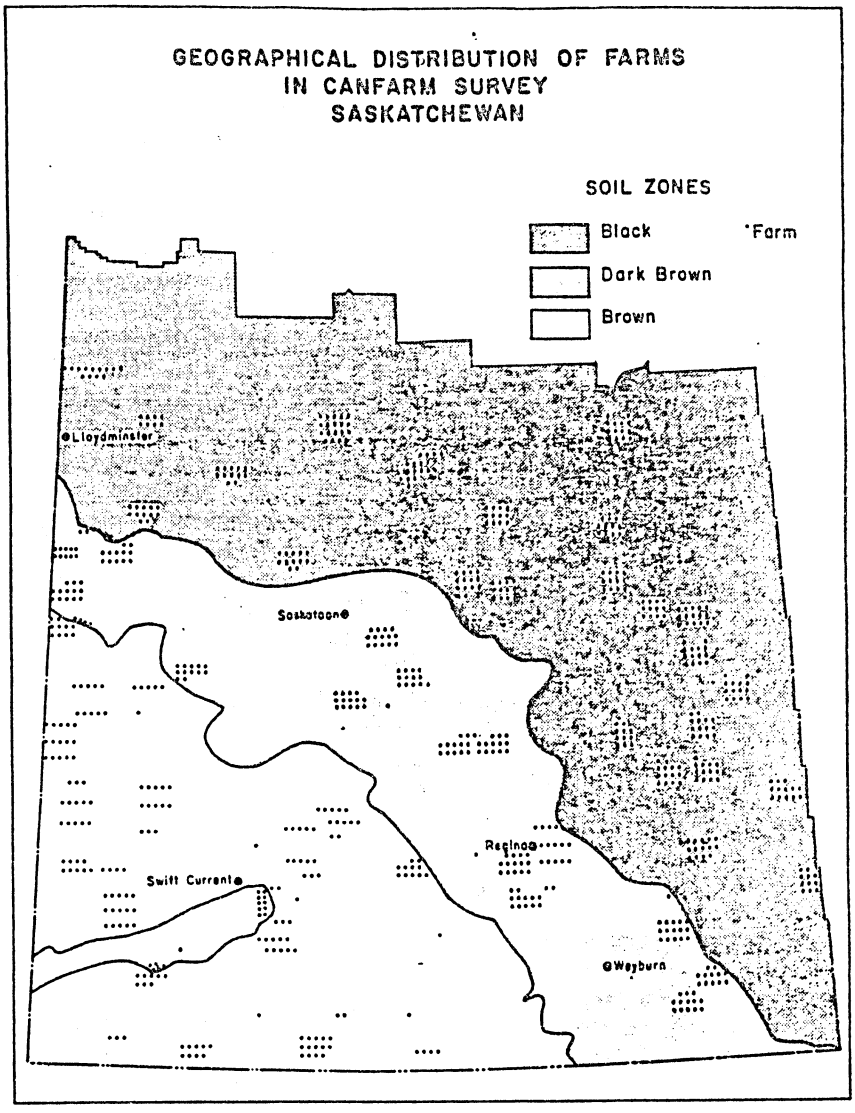
GEOGRAPHICAL DISTRIBUTION OF FARMS
IN CANFARM SURVEY
SASKATCHEWAN
SOIL ZONES
Black
Dark Brown
Brown
*Farm
Lloydminster
Saskatoon
Regina
Swift Current
Weyburn

TABLE 1

THE COEFFICIENTS (A_{1I}) AND (A_{2I})
OF THE FIRST TWO PRINCIPAL COMPONENTS

EIGEN-VALUE		Y_1^*	Y_2^*	Y_3^*	Y_4^*	Y_5^*	Y_6^*	Y_7^*	Y_8^*	Y_9^*	Y_{10}^*	Y_{11}^*
$\lambda_1 = 3.59$	Z_1	-0.49	-0.16	-0.17	-0.04	0.05	0.05	-0.45	-0.48	-0.34	-0.38	-0.05
$\lambda_2 = 1.43$	Z_2	-0.03	-0.27	-0.02	-0.61	0.60	0.26	0.18	0.11	-0.11	-0.11	-0.24

NOTE: Z_1 AND Z_2 DENOTES THE FIRST AND SECOND PRINCIPAL COMPONENT,

λ_1 AND λ_2 ARE EIGENVALUES, AND

Y_I^* DENOTES ORIGINAL VALUES Y_I RESCALED TO UNIT VARIANCE AND SHIFTED TO ZERO MEAN.

FIGURE 2

A Plot of the First Two Principal Components

Plot of Z_2: second principal component of the correlation matrix
Against Z_1: first principal component of the correlation matrix

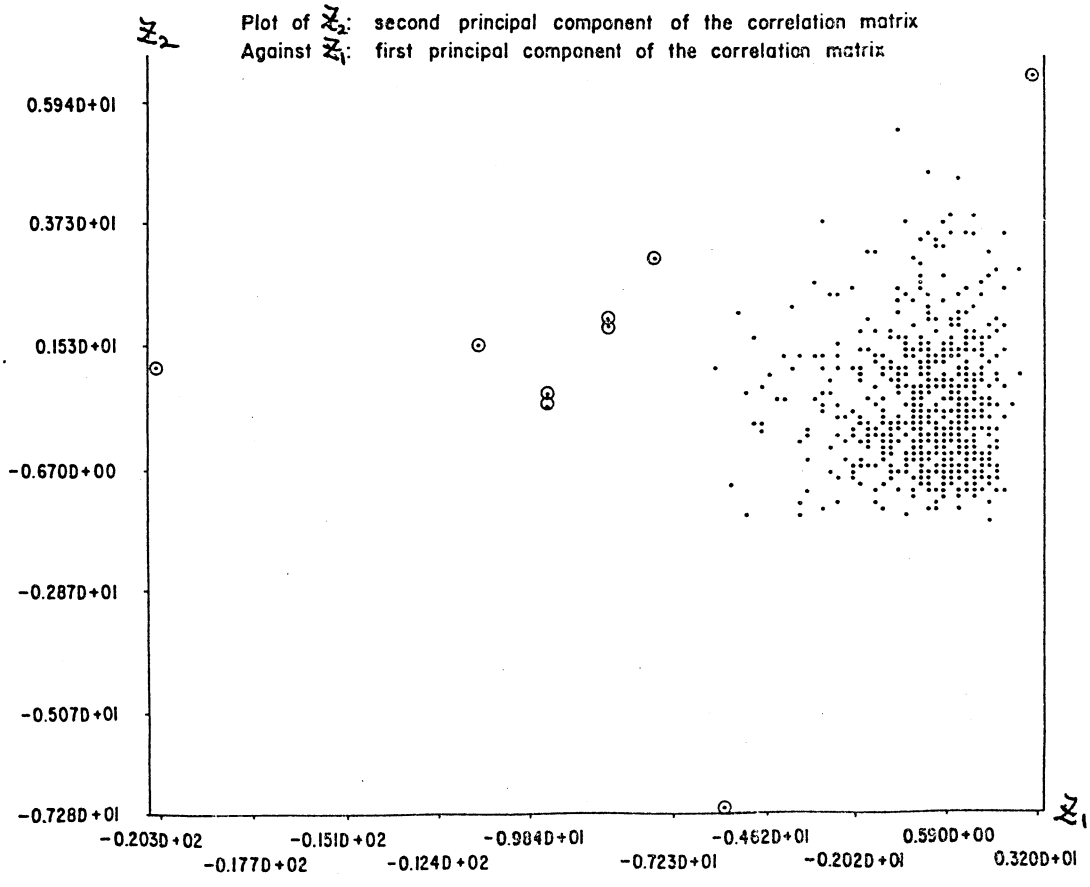

TABLE 2

OLS RESULTS FOR THE AVERAGE PRODUCTION COST MODEL

Dependent Variables \ Independent Variables	Const.	$1/S_{ij}$	FE_{ij}	$SLCW_{ij}$	SP_{ij}	CEI_{ij}	F	$\bar{R}^2$	N
1. APC_{ij}	0.041* (4.202)	10731.043* (42.996)	0.080* (4.794)	0.062* (4.960)	-0.053* (-4.584)	$-0.857^{(-4)}$* (-7.116)	380.168	0.739	670
2. APC_{ij}	0.073* (12.456)	5132.846* (23.521)	0.029* (2.873)	0.065* (6.597)	-0.057* (-8.233)	$-0.490^{(-4)}$* (-7.682)	135.600	0.505	661

Where N = Number of observations.
* = t value is significant at 1 per cent level.

FIGURE 3

COMPARISON OF AVERAGE COST CURVES