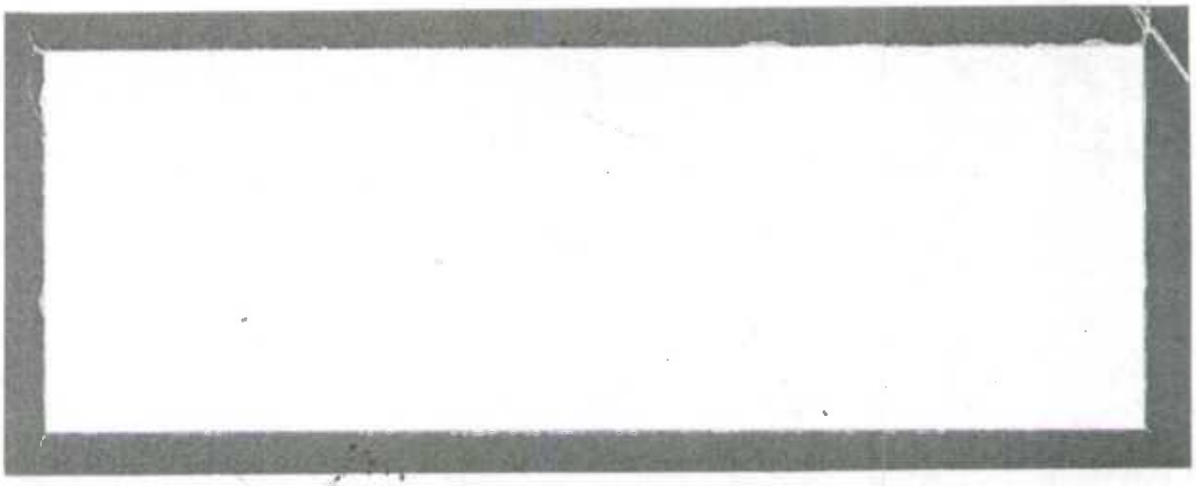


Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

WORKING PAPER NO. BSMD-90-005E/F

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-90-005E/F

DIRECTION DE LA MÉTHODOLOGIE

ANALYSIS ON GROUPING OF VARIABLES AND ON DETECTION OF QUESTIONABLE UNITS

by

France Bilocq and J.-M. Berthelot

March 1990

ANALYSIS ON GROUPING OF VARIABLES AND ON DETECTION OF QUESTIONABLE UNITS

by

France Bilocq and J.-M. Berthelot

March 1990

TABLE OF CONTENTS

1 INTRODUCTION

Editing in the general data collection and capture function is not an isolated step. It is part of an integrated process that comprises the following three stages :

1. First, natural groups of variables are identified, so that the application of editing procedures can be optimized;

2. Next, questionable units are identified by means of statistical techniques. A questionable unit is a unit with one or more variables whose values deviate significantly from the rest of the population;

3. Lastly, the questionable units are classified (score function) for confirmation and/or correction by follow-up procedures or for imputation by the general imputation system.

The purpose of this document is to present the results of empirical studies about the grouping of variables and the detection of questionable data. The score function is presently under study. The results will be documented later.

2 GROUPING METHOD

To facilitate the development of data editing methodology, data analysis techniques are used to design an edit scheme for quantitative variables from economic surveys. Factor analysis is a branch of multivariate analysis which concentrates on the internal relations of a set of variables. The grouping method described below is based on this theory.

The grouping method is a statistical tool used to identify natural groups of variables. The objective of this method is to partition the set of variables into two or more disjoint groups. Each of the groups identified contains variables that are highly correlated with each other. The results of the grouping method are then used to develop an editing scheme allowing only variables within the same group to be cross referenced in order to minimize the total number of cross editing rules.

2.1 Description of the grouping method

The grouping method divides up a set of variables on the basis of their correlation matrix. Each group is identified in such a way as to maximize the variance accounted for by the first principal component and to ensure that the variance accounted for by the second principal component is not too high. These two constraints are controlled by means of parameters.

Initially, the method treats all variables as belonging to the same group. Then the following steps are repeated until all the conditions are met.

1. The first step consists of performing a principal component analysis on the correlation matrix of each group independently.

2. A given group is subdivided if the proportion of variance accounted for by its first principal component is insufficient or the proportion of variance explained by the second principal component is too high. Otherwise, a group is not split.

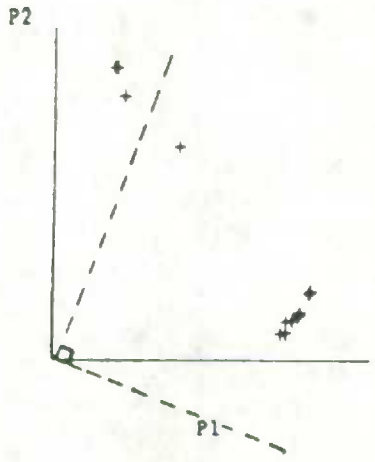
3. A group is divided by using a variant of the orthoblique rotation proposed by H.H. Harman (1976) [2] and W. Harris and F. Keiser (1964) [3]. An oblique solution may be obtained from an orthogonal rotation combined with a transformation defined by a positive definite matrix. An oblique solution implies that the reference axes are not orthogonal.

   One of the basic principles of factor analysis is to always search for a simple solution. By simple solution it is meant a solution for which each variable is represented by a small number of axes (preferably only one). However, such a solution is not always found automatically by the factor analysis of a correlation matrix. This is why rotation of axes is used.

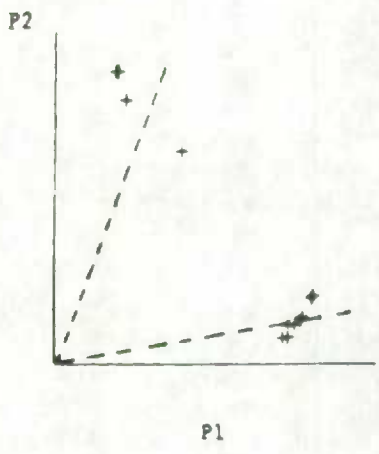
   Principal component analysis provides one of the infinite number of possible solutions for the reference axes. Rotation allows pivoting these axes so as to obtain a better representation of the variables on them. These steps help to find a solution that explains the variability of a group of variables more adequately.

   Why look for a rotation which is orthoblique rather than orthogonal ? By forcing the rotation to be orthogonal, the same constraint is imposed on the axes. To get a clear picture and to ensure that each variable is well represented by one axis only, it is often required that the axes be oblique. For example, a correlation exists between the variables of the non-manufacturing and the variables of the manufacturing products (Census Of Manufactures, COM). However this link is not strong enough to justify grouping these two categories of variables in such a way that they are well represented by only one axis. If an orthogonal rotation is imposed, the solution obtained will clearly show that one group is at a disadvantage to the other. If an oblique solution is allowed, it is then possible to clearly represent both groups simultaneously with different axes. In order to visualize this situation, the plane of the first two principal components is presented for an orthogonal and an oblique solution in the following diagrams.

Orthogonal rotation



Oblique rotation



Note: Principal component analysis extracted from the Census Of Manufactures data.

The oblique solution is easier to interpret. Using such a solution allows for a more detailed classification of variables.

The number of groups obtained by the grouping method is controlled by the use of parameters specifying the proportion of variance explained by the first two principal components. In fact, if the parameter for the variance explained by the first principal component is increased (or the parameter for the variance explained by the second principal component is decreased), then the number of groups obtained will be higher.

To split a group in two, the correlations between the variables and the first two principal components obtained by the orthoblique rotation are used. Each variable is paired with the axis with which it is most correlated. Since the orthoblique rotation provides a solution with axes that are not orthogonal, the method provides the correlation coefficient between the groups.

4. An iterative process reassigns variables (if necessary) in the groups so as to maximize the variance accounted for by the first principal component of each group.

5. Back to step 1 : The process ends when all groups meet the specified criteria for the proportion of variance accounted for by the first two principal components. The algebraic details of the grouping method are provided in appendix 1.

2.2 Correlation Coefficients

The grouping method uses a correlation matrix as input. Various types of correlation coefficients may be used.

Pearson's correlation coefficient measures the degree of linear association between two variables. It is easy to interpret and also very quick to compute. However, it is influenced by large values. In economic surveys, this characteristic tends to lead to overestimation of the correlation between variables.

Spearman's and Kendall's correlation coefficients are non-parametric coefficients. They measure a degree of association which is not necessarily linear. For example, a strictly increasing function between two variables will produce Spearman's and Kendall's coefficients with a value of one. These two coefficients are harder to interpret and slower and more expensive to compute.

As a general rule, linear relationships are found between variables in an economic survey and since Pearson's coefficient is better known, easier to interpret and faster to calculate, it was used for the purposes of the empirical studies. However, as discussed later, transformations are needed to minimize the impact of large values on the Pearson correlation coefficient.

2.3 Empirical Studies

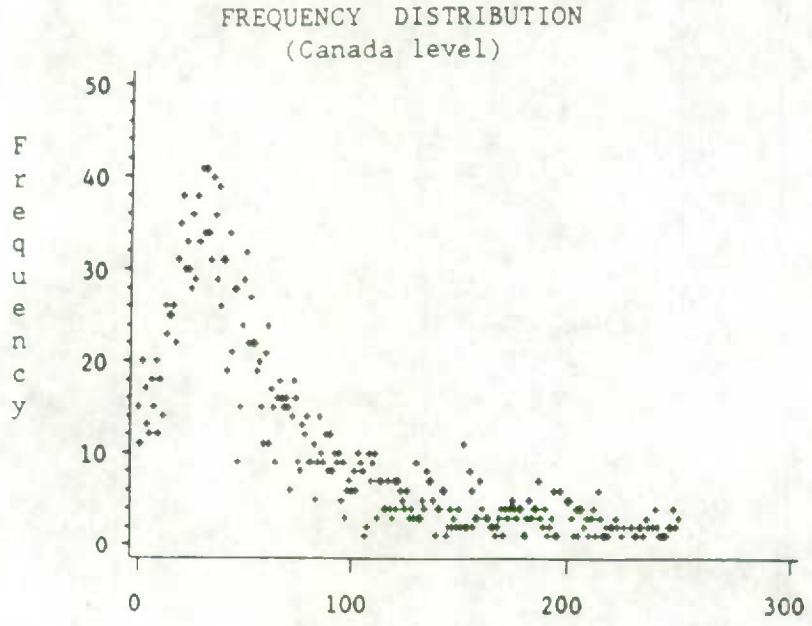
2.3.1 Data used

Empirical studies were conducted in order to analyse the behaviour of the grouping method on real data. For these studies, the SAS procedure VARCLUS [6] has been used for the grouping of the variables. This procedure uses a technique similar to the one described earlier.

Final data from the 1984 and 1985 Census Of Manufactures (COM) long questionnaires were used for this purpose. Final data have been edited, corrected and imputed; i.e., data used for publications. There are several versions of the long questionnaire, which vary according to the standard industrial classification. Only standard financial variables were analysed. These are the variables found systematically in every version of the long questionnaire. A sample of the standard questionnaire is given in Appendix 2.

2.3.2 Preliminary analysis

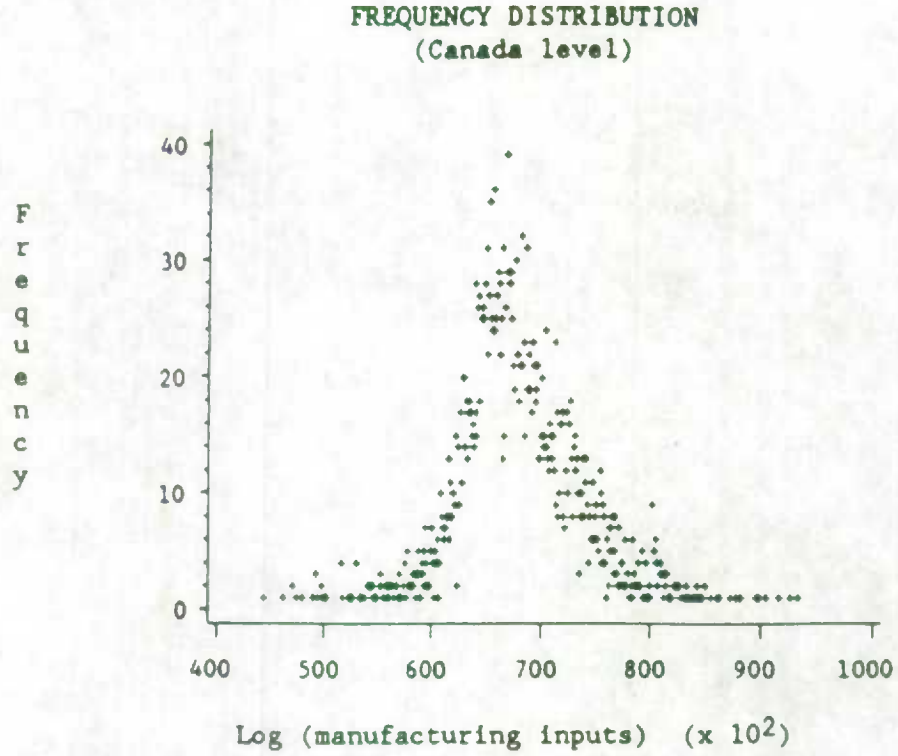
Before going further into the application of the grouping method, the frequency distribution of the variables under study was analysed. The graph below represents the frequency distribution (at Canada level) of one of the financial variables studied.



The variable has an asymmetrical frequency distribution very close to that of a lognormal. There are many units with low values and very few with large values. All the financial variables studied have similar frequency distribution. Analysis at lower levels (provinces and standard industrial classification) yielded similar results.

The grouping method uses Pearson's correlation coefficient which is influenced by large values. Accordingly, data were transformed to obtain a more symmetrical distribution and thus reduce the impact of large values on the correlation coefficient. The transformation used was the logarithm. The following graph represents the frequency distribution of the transformed variable.

A distinct improvement in the symmetry of the distribution can be observed. When the grouping method is applied, Pearson's correlation coefficient will be calculated from the transformed variables (log). The correlations will consequently be influenced less by large values and thus be more conservative. The results of more studies of correlation coefficients are presented in Appendix 3.



2.3.3 Results (grouping method)

The grouping method was applied to the 1984 and 1985 Census Of Manufactures (COM) data. Two studies were conducted.

First study

The first study was conducted at the Canada level for the 1984 and 1985 COMs, and entailed classifying the 73 financial variables from the long questionnaire, without taking into account the questionnaire's structure.
The results obtained for the 1985 COM are as follows :

- Thirty groups were identified, ranging in size from one to nine variables
- The groups are consistent in that the variables in a group are logically interrelated;
- The results were confirmed by the COM subject matter officers

The results obtained for the 1984 COM are as follows :

- Twenty-eight groups were identified, two less than for the 1985 COM;

- An analysis of the results shows that 24 of the 28 groups identified for the 1984 COM are exactly the same as those identified for the 1985 COM. The other four originate from separating or combining some of the groups identified for the 1985 COM; On the basis of the findings, it was assumed that historical data (1984 COM) and current data (1985 COM) behaved similarly.

  From this hypothesis, the grouping method was applied simultaneously to the 146 variables of the two censuses. The results were conclusive. Thirty-six groups were obtained. One may note that the majority of the groups were the same except twice as big. The same variables from the two censuses were paired, which on the whole confirms the similarity of the data from one cycle to the next of the Census Of Manufactures. A list of groups identified by the method in the above three experiments is given in Appendix 4.

Second study

The purpose of the second study was to evaluate the behaviour of the method when applied at various levels including Canada, the provinces and standard industrial classification. This time, the variables to be studied were chosen on the basis of the questionnaire structure. The twelve most important variables were used. These are divided into four general classes :

i) Opening inventory
ii) Closing inventory
iii) Inputs
iv) Outputs

These are in turn subdivided into three components :

a. Manufacturing products
b. Non-manufacturing products
c. Total manufacturing and non-manufacturing products.

Variables 1a, 1b, 2a and 2b were created by adding up the appropriate components from the original questionnaire.

The first step to perform was to classify the variables at Canada level. The results obtained are presented in the following table. The number in each box corresponds to the group in which the variable was classified.

RESULTS AT CANADA LEVEL

VARIABLES	Opening inventory	Closing inventory	Inputs	Outputs
Manufacturing Products	1	1	2	2
Non-manufacturing products	3	3	4	4
Total	1	1	2	2

Four groups were identified. It is known that opening and closing inventories are closely related and that a similar link exists between inputs and outputs variables. It is also true that the manufacturing component is the one that contributes most to the total. The groups are thus representative of the specific characteristics observed among the data.

The grouping method was then applied by province. For each of the ten provinces, the set of twelve variables was divided into four parts. Exactly the same groups were obtained as for the classification at Canada level.

Lastly, the method was applied at the two digit standard industrial classification level (SIC2). Twelve SIC2s were analysed. For six of them, the results were the same as before. For the other six SIC2s, however, three groups were identified instead of four. Analysing the situation a little more closely, it has been found that in all six cases, the third group corresponded to a combination of two of the four basic groups : either a combination of groups one and two or a combination of groups three and four. These results are consistent with the characteristics stated above regarding manufacturing and non-manufacturing components.

NOTE : By changing the value of the parameters (variance accounted for by the two first principal components), the four basic groups were obtained for the twelve SIC2s.

2.4 Conclusions

The results of the empirical studies are conclusive. The grouping method extracts the correlation pattern easily from a set of variables. All the grouped variables are logically interrelated and the results are confirmed by COM subject matter officers.

Given that suitable software exists in the SAS software package [6], it is easy to implement. Moreover, it is easy to use and interpret as well as being economical. The transformation (log) of the data and the calculation of Pearson's correlation coefficients are operations that can be performed quickly. Performing multivariate analysis on the correlation matrix also leads to savings, since the work is done only on a small set of data. It is important to remember that the aim of the grouping method is to group interrelated variables together for edit purposes. Once the groups have been identified, cross-editing can be confined to the relations between the variables in one group while ensuring a high degree of consistency for the record. Edits involving variables from different groups will be carried out only if they are necessary to ensure a higher degree of consistency for a record. The grouping method provides coefficients of correlation between groups. This information could, in some cases, guide the choice of inter-group edits.

The grouping method is a tool that combines statistical theory and experience with data. One of the objectives of statistical theory is to provide a scientific law or a mathematical model to explain the behaviour of the data. A prior knowledge of the behaviour of the variables under study can help to verify the consistency of the groups identified by the method. Furthermore, certain unknown links between variables may be revealed by applying this method. However, if the behaviour of the variables under study is unknown, or if experimental results are desired, then the method groups the variables empirically and objectively. This method thus enables us to place the data edit process in a theoretical framework without forgetting its intuitive nature.

3 DETECTION METHOD

The purpose of this section is to present the results of the empirical studies designed to evaluate the behaviour of two bivariate methods for detection of questionable data, namely simple linear regression and the tolerance method. Multivariate detection methods will be the subject of future research.

3.1 Definition

A bivariate detection method is used to model the links between two variables from the same or different survey cycle(s) using data from previous cycle(s) to help to identify suspicious behaviour of these two variables in the current cycle.

To determine whether or not a detection method was effective, the following evaluation criteria were used :

- To make the behaviour of a bivariate detection method easier to visualize, a graph representing the values of one variable versus the values of the other variable was plotted. The acceptance interval obtained by the method was then indicated on the graph. The rejection bounds were expected to follow the shape of the data as closely as possible.

- Units with large values are very important. Sometimes they behave differently from the rest. A detection method should model adequately the behaviour of these units.

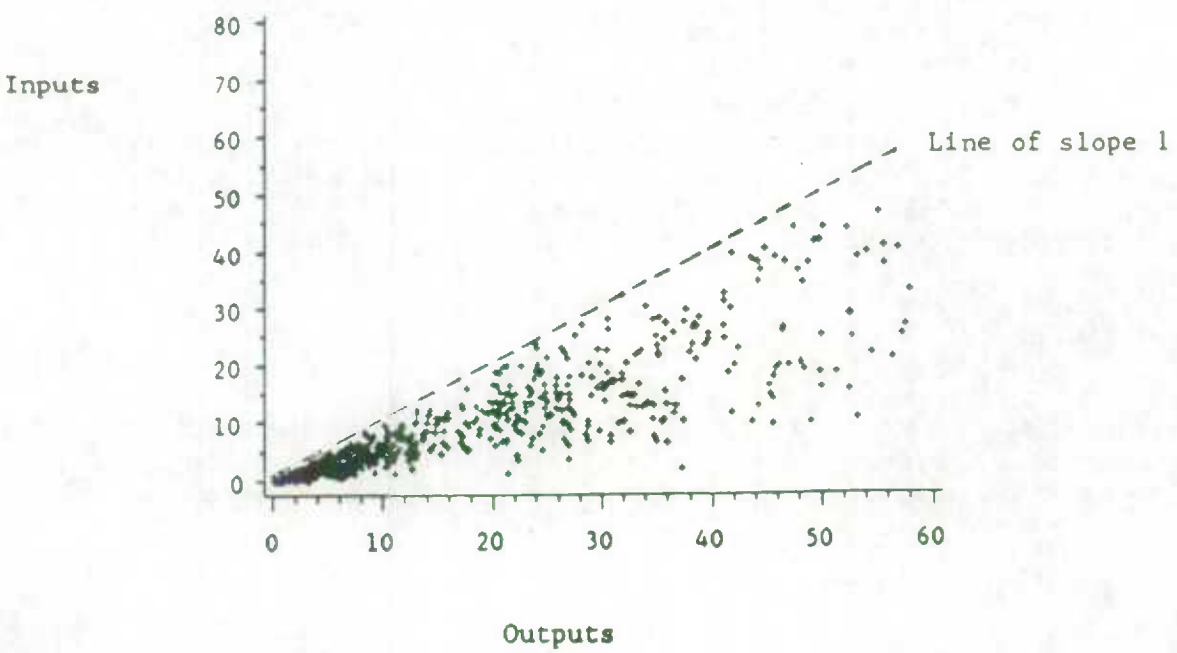
- Variables are not always reported or captured in the same order. A detection method should be independent of the order in which variables are given.

3.2 Empirical Studies : data used

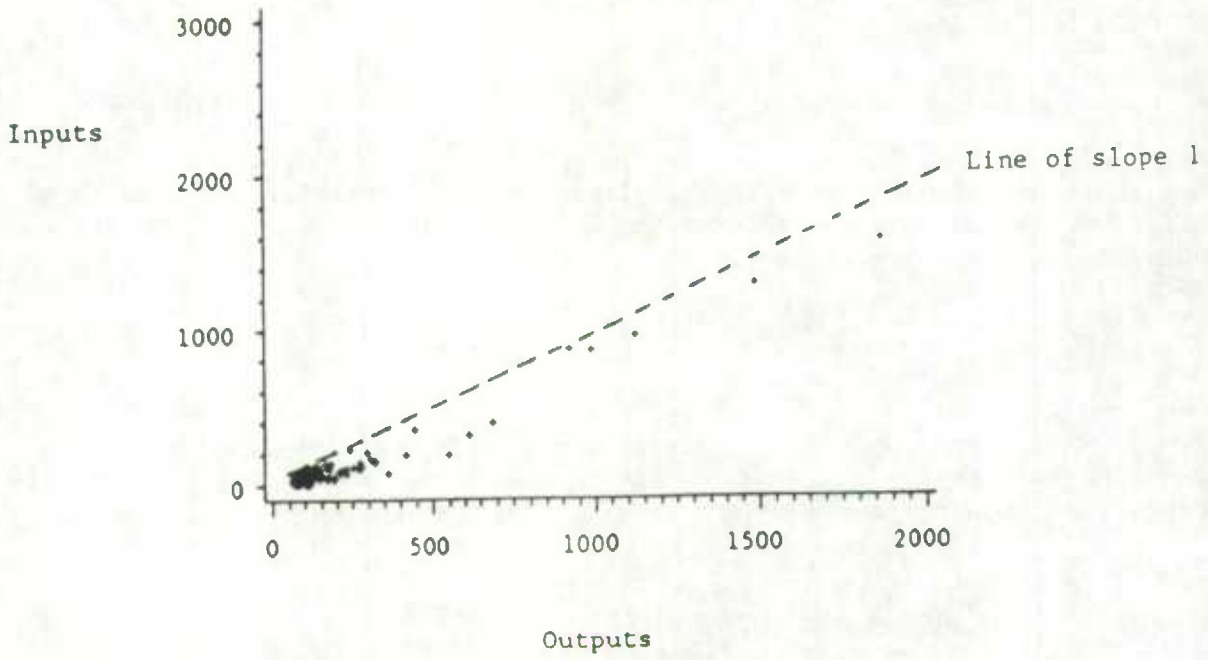
The final data from long questionnaires for the 1985 Census Of Manufactures were used for the studies of both detection methods. However, only the 2,887 units for the province of Québec were used. The two variables used were manufacturing inputs and outputs.

The graph that follows represents the variable "manufacturing inputs" as a function of the variable "manufacturing outputs". So that the shape of the graph could be visualized better, it was divided into two parts. The first graph represents all units whose manufacturing outputs variable is less than or equal to sixty million dollars (95% of units) and the second represents the remainder of the units.

OUTPUTS vs INPUTS
(Province of Québec)
(units with outputs <= $60 million)
(Graph in million of $)



INPUTS vs OUTPUTS
(Province of Québec)
(units with outputs > $60 million)
(Graph in million of $)



If one considers the above graphs, it can be noted that variability increases with the value of the output variable. The shape of the data resembles a funnel. Furthermore, there is a boundary in the upper part of the first chart, above which there are very few units. This is due to a natural constraint, namely that apart from exceptional cases, inputs are always smaller than outputs. This constraint corresponds to a line with a slope of one.

3.3 Simple Linear Regression

3.3.1 Description of the method

The first method tested was simple linear regression. Suspicious links between two variables are identified by means of a 95 % confidence interval calculated around the regression line. All units located outside the confidence interval are considered questionable.

It is well known that regression is a statistical method that is influenced by extreme values. Thus the confidence interval around a regression line goes further away from the line as values increase. Hence because of the properties identified, the regression was expected to be an appropriate detection method, since the confidence interval will tend to follow the funnel shape of the data.

Final data for the 1985 COM (Québec) were modelled and the results were applied to the same set of data. When the edit procedures are applied, data from a previous cycle will be modelled and the results will be applied to current cycle data.The adjusted regression model is of the following type :

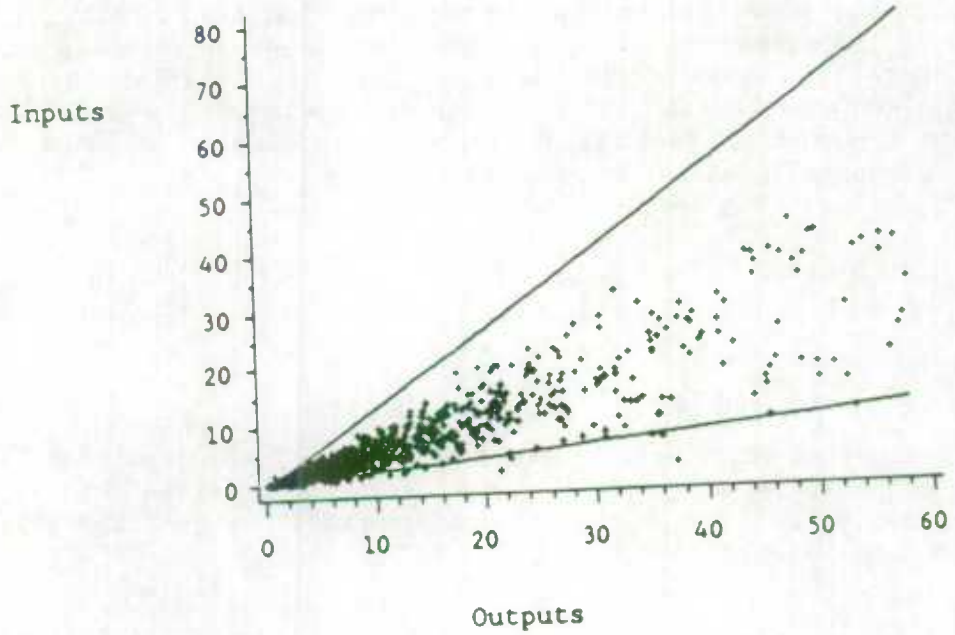
$$Log(inputs) = \alpha + \beta log(outputs) + e$$

The log function was used so as to stabilize residual variability. Details of the model and the residual analysis are given in Appendix 5. The results confirm the validity of the adjusted model.

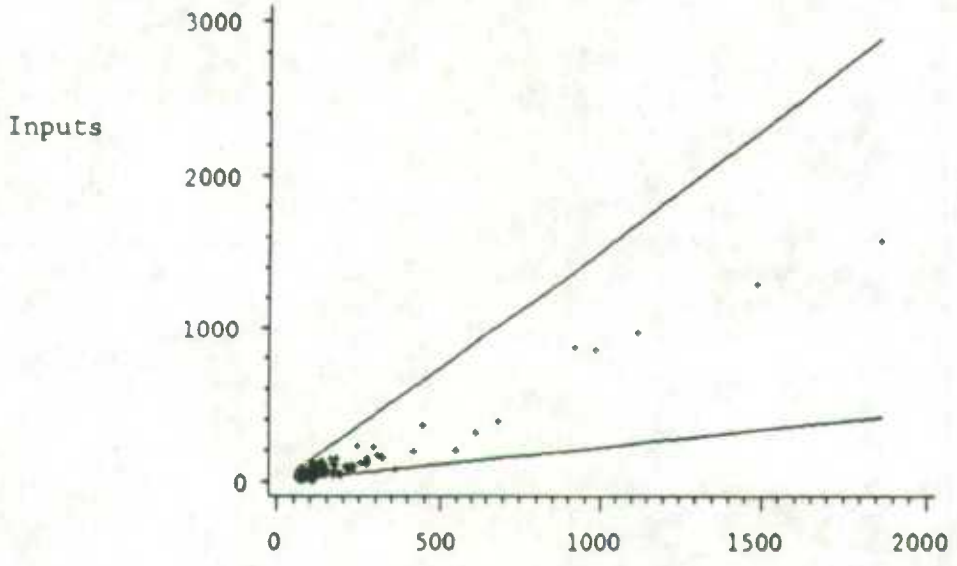
3.3.2 Results

The speed at which the confidence interval deviates from the regression line was under-estimated. The graphs that follow represent the confidence interval of the regression line projected on the original (untransformed) data plane.

REGRESSION LINE CONFIDENCE INTERVAL
(units with outputs <= $60 million)
(Graph in million of $)



REGRESSION LINE CONFIDENCE INTERVAL
(units with outputs > $60 million)
(Graph in millions of $)



The application of the linear regression method led to rejection of 108 units. The confidence interval does not follow the shape of the data very well. In fact, the greater the values the greater the deviation. The majority of units rejected by this method are thus units for which both variables have small values.

3.3.3 Conclusions

Simple linear regression is too strongly influenced by extreme values. The confidence interval does not follow the shape of the data. The width of the interval increases (too fast) with the size of the values. Consequently, the regression method is not strict enough for large values.

Moreover, the results differ depending on the order in which variables are processed. Regression model $x = \alpha + \beta y$ does not identify the same questionable units as model $y = \alpha + \beta x$. The results of the linear regression detection method thus depend on the order of the variables.

Hence simple linear regression is not an appropriate detection method, since it does not meet the established criteria. Other avenues, however, may be explored; for example, non-parametric regression and regression using dummy variables. Alternatively, tools other than the confidence interval could be used to identify questionable units. These avenues will be explored later if time and budget allow.

3.4 Tolerance Method

This method identifies units in which the trend (relationship) between two variables differs significantly from the corresponding overall trend of the other units. It attempts to model the trend between these two variables. The relationship can be defined in one of two ways : either as a ratio between two variables in the same survey cycle or as a relative difference between a variable in a current cycle and the same variable in a previous cycle.

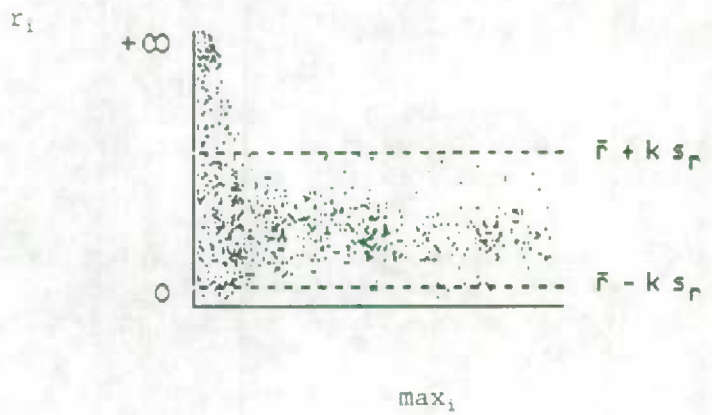
The monthly manufactures survey (CSIO) is now successfully using the tolerance method for the editing system in regional offices. The method is used to control the relative difference between a variable in the current month and the corresponding variable in the previous month. Since the method has demonstrated its effectiveness for this purpose, the goal of the present study is thus to verify the method's effectiveness in controlling the ratio between two variables in the same cycle.

3.4.1 Description of the method

The purpose of this section is to provide the theory underlying the tolerance method [4], which seeks to model the ratio between two variables. The shape of the graph for the ratio between two typical economic variables (r_i) versus the maximum between the two variables is represented in the following figure.

$r_i = x_i/y_i$ $\qquad$ $max_i = max\ (x_i, y_i)$

GRAPH OF THE RATIO vs THE MAXIMUM OF THE TWO VARIABLES



The graph is not symmetrical, it shows compression toward zero. Moreover, there is greater variability in ratios for small units than for large ones.

If bounds such as $\bar{r} + ks_r$ (where $\bar{r}$ is the average ratio and s_r is the standard deviation of ratios) are applied, the edit rejects would have over-representated small units and under-representated large units. This is known as the size masking effect, and is due to the difference in variability between small and large units. In certain circumstances, this method may not allow the detection of questionable units in the lower part of the graph (negative lower bound) [8].

To correct this problem, the ratio r_i must be transformed in such a way that questionable units can be detected at both extremes.

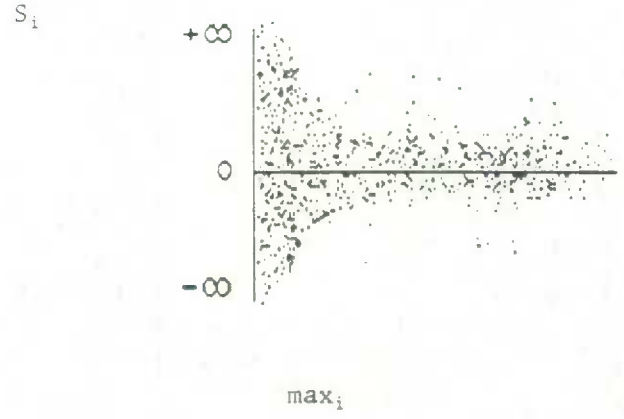
The graph must be made symmetrical in relation to zero. A possible transformation is defined as follows :

$S_i = 1 - r_M/r_i$ if $0 < r_i \leq r_M$

$r_i/r_M - 1$ if $r_i > r_M$

where r_M is the median of the r_i

GRAPH OF S_i vs THE MAXIMUM OF THE TWO VARIABLES

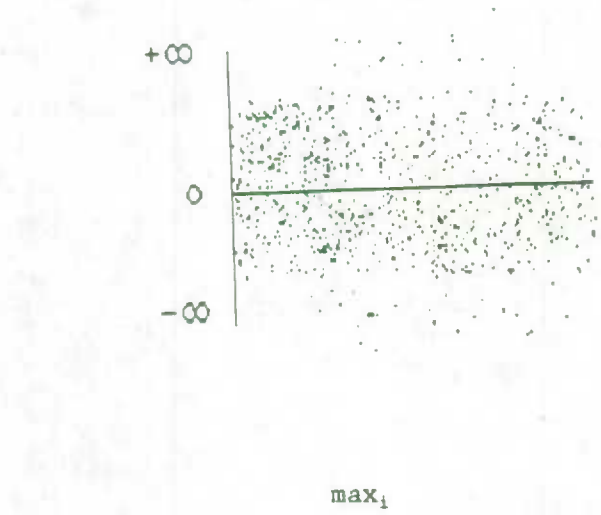


Even with this transformation, the size masking effect is still present. The following transformation is performed in order to take into account the value of the data :

$E_i = S_i * [Max_i]^u$, where $0 < u \leqslant 1$.

The values of E_i represent the effects and the exponent u determines the importance of the value of the data. This transformation gives greater importance to a small variation in a large record than to a large variation in a small record.

GRAPH OF E_i AS A FUNCTION OF THE MAXIMUM OF THE TWO VARIABLES



If the first quartile, the median and the third quartile of E_i are designated as E_{Q1}, E_M and E_{Q3} respectively, then an acceptance interval can be defined as follows :

Lower bound = $E_M - c_l * (E_M - E_{Q1})$

Upper bound = $E_M + c_u * (E_{Q3} - E_M)$

The parameter c_l and c_u determine the breadth of the acceptance interval and the parameter u determines the curvature of the two bounds. All units whose corresponding value of E_i falls outside these bounds are defined as questionable.

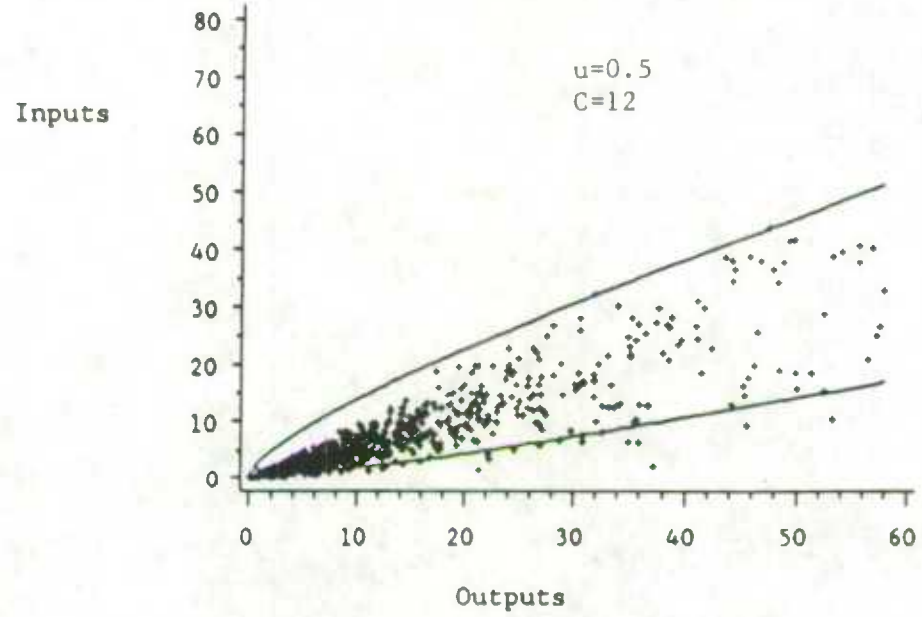
3.4.2 Results

This study used the same set of data (Québec 1985) and the same variables (manufacturing inputs and outputs) as the regression method.

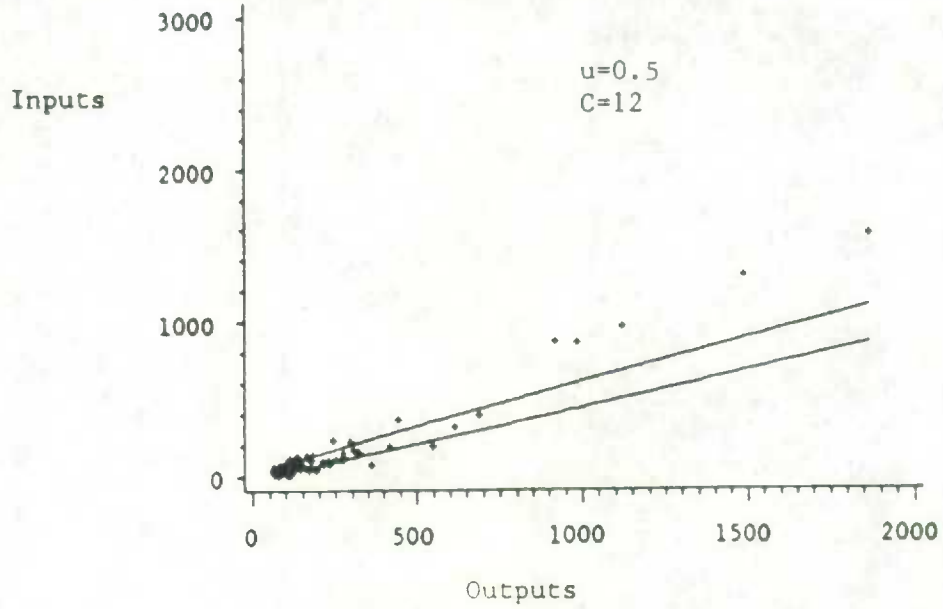
The ratio studied was : $\frac{\text{manufacturing inputs}}{\text{manufacturing outputs}}$

Final data from the 1985 COM (Québec) were modelled and the results were applied to the same set of data. When edit procedures are applied, data from a previous cycle are modelled and the results applied to current cycle data. The following graphs (in millions of dollars) represent the acceptance interval obtained by the tolerance method in the original data plane.

TOLERANCE METHOD ACCEPTANCE INTERVAL
(units with outputs <= $60 million)



TOLERANCE METHOD ACCEPTANCE INTERVAL
(units with outputs > $60 million)

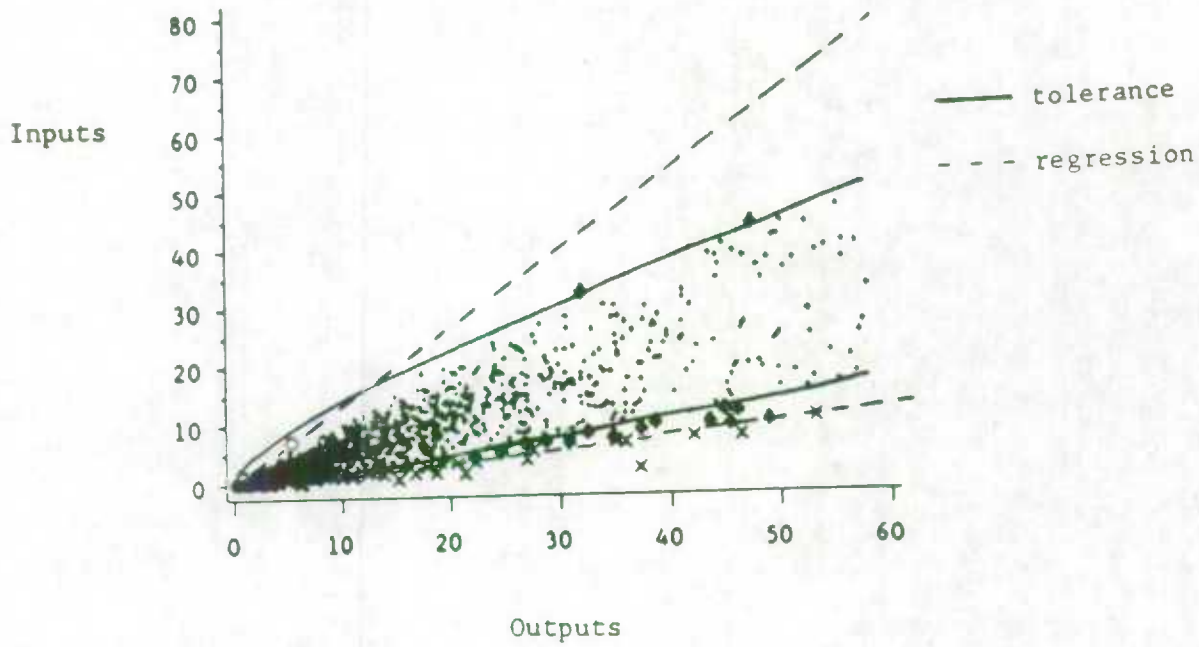


Using the tolerance method, 133 units were rejected. The two bounds closely follow the shape of the data and even seem to draw in closer for large values. The method thus tends to be stricter for these values. Very few units located in the upper part of the graph were rejected, because of the natural constraint identified earlier, i.e., that with few exceptions, the manufacturing inputs are always smaller than manufacturing outputs.

3.4.3 Comparison

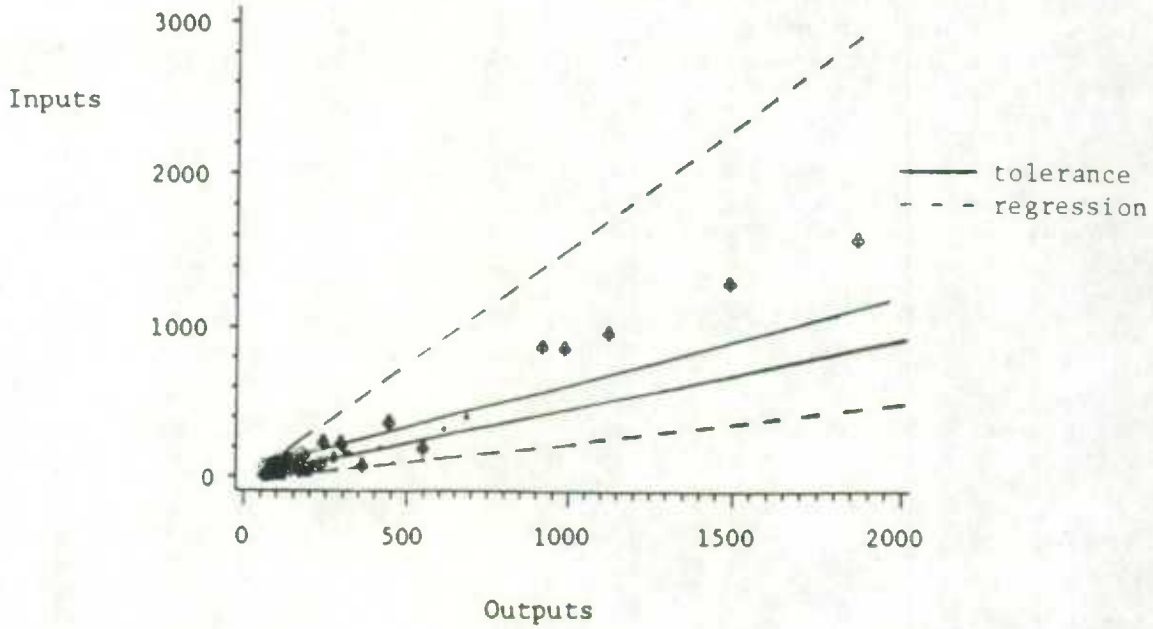
Now let us compare the results obtained by the two methods. The graphs that follow represent the results of the two edit methods simultaneously.

REGRESSION AND TOLERANCE METHOD RESULTS
(units with outputs <= $60 million)
(Graph in millions of $)



Legend : + unrejected units
O units rejected only by the regression method
◆ units rejected only by the tolerance method
x units rejected by both methods

REGRESSION AND TOLERANCE METHOD RESULTS
(units with outputs >$60 million)
(Graph in million of $)



Legend : + unrejected units
⊕ units rejected only by the tolerance method
x units rejected by both methods

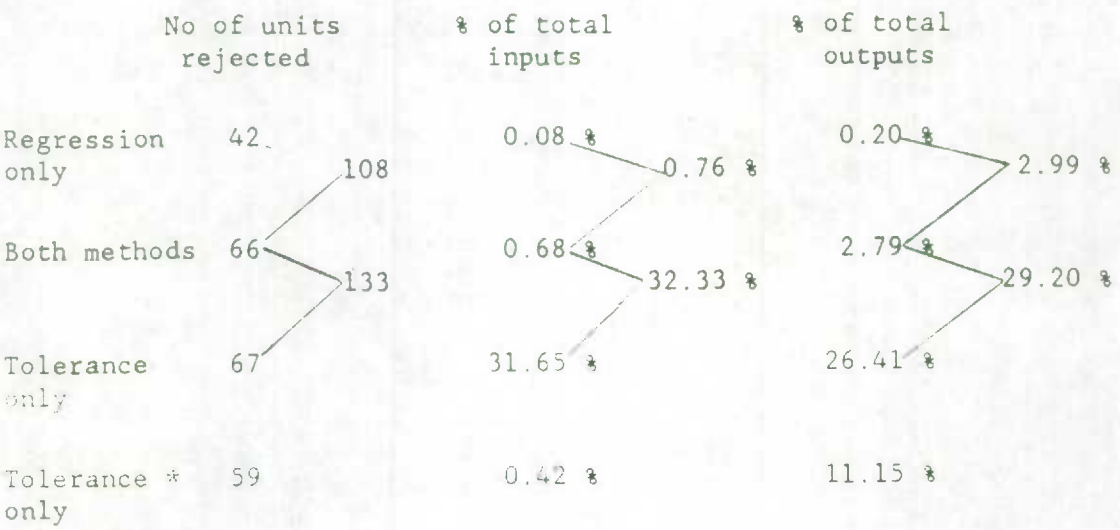
The gap between the two intervals widens as outputs increase. The greater the values the more the regression bounds deviate from the data, whereas the opposite phenomenon can be observed for the tolerance method.

The units rejected by the two methods are not all the same. The table that follows shows the number of units rejected by the regression method only, the number of units rejected by both methods and lastly the number of units rejected only by the tolerance method.

	No of units rejected by		
Regression only	42		
		108	(regression)
Both methods	66		
		133	(tolerance)
Tolerance only	67		

Referring to the previous graphs, it can be seen that the regression method rejects units for which the two variables under study have smaller values. Statistics have been compiled in order to assess the impact of each method on the type of units rejected. A proportion has been calculated for the two variables and the two methods under study. It is the corresponding $ value of a variable for units rejected by a particular method divided by the total $ value of this variable (all the units).

Total manufacturing inputs : $ 29 million
Total manufacturing outputs : $ 51 million



* Recalculated omitting the eight units with largest value of outputs.

The 42 units rejected only by regression correspond to .08% of total inputs and .2% of total outputs. These are very small proportions when compared to the units rejected only by the tolerance method, which correspond to 31.65% of total manufacturing inputs and 26.41% of total manufacturing outputs.

The results obtained by the tolerance method may seem "inflated", since it is obvious when looking at the earlier graphs that this method identifies units in which the values of both variables are extremely large. As a result, the proportions were recalculated, with the eight units for the largest output values omitted. The results remain conclusive notwithstanding. The remaining 59 units account for 10% and 11% respectively of the totals of manufacturing inputs and outputs. The units identified as questionable by the tolerance method are much more significant in terms of impact on totals than those identified by the regression method.

3.4.4 Conclusions

The analysis of the tolerance method leads to the following conclusions. The upper and lower bounds of the tolerance method closely follow the outline of the point cloud and adequately model the behaviour of units with large values. Moreover, the method is symmetric in that the same questionable units will be identified whether working on the ratio x/y or the ratio y/x.

This method has other interesting properties. The distance between the two bounds and the form of the curve are controlled by means of parameters. The method thus offers the possibility of controlling the number of rejections up to a certain point. Each bound may also be adjusted separately. This offers the power to model units located in the upper part of the graph differently from the rest. The tolerance method is also easy to implement. All that is needed is to use the data from an earlier cycle to calculate acceptance bounds according to the desired level. When data are edited, the ratio between the two variables to be edited is calculated and then compared to the predetermined bounds.

The tolerance method meets the established criteria. Moreover, it identifies a minimum number of units as questionable. This implies that the number of confirmations/corrections to be made can be minimal, while data quality is maintained.

For example, the statistics that follow were collected from a parallel study made for the monthly manufactures survey (CSIO). The purpose of this study was to compare the old (manual) detection method with the tolerance method. The results were as follows :

MONTHLY MANUFACTURES SURVEY (CSIO)
Percentage of rejections
(Questionable units requiring confirmations/corrections)

	manual method	tolerance method
Halifax	33.2 %	15.8 %
Winnipeg	41.9 %	14.1 %

By using the tolerance method, the number of rejections has been cut by half. This implies a cost saving, since the number of confirmations/ corrections to be performed is reduced. However, what did happen to the data quality ? The answer to this question is the following : out of the 131 units actually corrected by interviewers using the manual method (old), 129 were identified by the tolerance method.

With the tolerance method, the number of questionable units decreases hence reducing editing costs. Since the vast majority of units that really need to be corrected are identified then the same degree of data quality is maintained. Thus there is a great opportunity for savings in costs and resources allocated to the editing process by using a method such as the tolerance method.

4 CONCLUSIONS

The purpose of this section is to recapitulate the significant results emerging in the course of our research. The aim in applying the grouping method is to identify groups of variables in order to maximize the effectiveness of the editing process. The SAS procedure is not the only method in existence, but the results obtained are conclusive. The grouping method easily extracts the correlation pattern from a set of variables and produces groups in which all variables are logically interrelated. The tolerance method performs well. It meets all the established criteria. It can be applied either to longitudinal detection (from one cycle of a survey to the next) or to detection within a single survey cycle. Moreover, it helps to reduce the costs associated with the data editing process.

5 FUTURE WORK

At some future point, various tasks will have to be performed in order to complete the research that has begun.

Development

After studying and developing a set of principles applicable to data editing, the most important task is to provide the support needed by the development team in order to prepare the production system for the generalized data collection and capture function. An equally important task is to complete the development of the score function. This is part of the integrated edit and correction process. Its purpose is to classify questionable units as required for follow-up procedures. This document is devoted to the processing of quantitative variables. An approach for qualitative variables within the framework of the generalized data collection and capture function might be developed.

Simulation

Several principles were put forward in the development of the generalized system. To put these principles into practice, the entire capture, edit and imputation process will be simulated using survey data. The results obtained will be evaluated in terms of costs and data quality.

Grouping method

In the empirical studies of the variable grouping method, Census Of Manufactures data have been used. The behaviour of the method should be studied using data from another survey.

Multivariate detection method(s)

The grouping method is used to identify groups of variables for editing. At present, only bivariate methods of editing have been studied. It would be interesting to be able to use a detection method that takes into account all the variables in a cluster. Since multivariate detection is a little-known subject, further research in that area must be conducted.

6 REFERENCES

[1] Berthelot, J.-M. (1989), General approach for the data edit and correction module, Statistics Canada, Business survey methods division, technical report.

[2] Harman, H.H. (1976), Modern factor analysis, 3rd ed, University of Chicago Press, Ill. (487 p.)

[3] Harris, W. and Keiser, F. (1964), Oblique factor analystic solutions by orthogonal transformations, Psychometrika 29, no 24, p. 347-362.

[4] Hidiroglou, M.A. and Berthelot J.-M. (1986), Statistical editing and imputation for periodic business surveys, Survey methodology 12, p. 73-83.

[5] Holzinger, K.J. and Harman, H.H. (1942), Factor analysis, University of Chicago Press, Ill. Chicago, Ill., (417 p.).

[6] SAS institue, (1985), Sas user's guide : statistics, Version 5, North Carolina.

[7] Thorndike, A.M., (1978), Correlation procedures for research, Gardner Press, New York.

[8] Wilkinson, R.G. (1982), Outlier identification technique designed for the Business Finance Annual Survey, Statistics Canada technical report.

APPENDICES

APPENDIX I: Theoretical description of the grouping method

This appendix presents an algebraic description of the principal components analysis and the orthoblique rotation of the grouping method.

1.1. Notation

I : Identity matrix

R : Matrix of correlations between variables

Q : Matrix of eigenvectors of the matrix R

L : Diagonal matrix containing eigenvalues of matrix R

A : Matrix of coefficients a_{ij} of the principal components analysis (factor loadings)

C_j : j^{th} principal component

T : Orthogonal rotation matrix

S : Matrix of correlations between variables and axes

F : Matrix of correlations between axes

P : Matrix of coefficients (factor loadings) obtained after transformation of the matrix A

D : Positive definite matrix

v : Number of variables

1.2. Principal component analysis (PCA)

PCA of the matrix of correlations of each group separately.

By matrix theory, a matrix A can be found such that $R = AA'$

By PCA theory, $R = Q\Lambda Q'$ (correlation matrix broken down into eigenvalues and eigenvectors)

If we define : $A = Q\Lambda^{\frac{1}{2}}$

We find that : $AA' = Q\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}Q'$

$= Q\Lambda Q'$

$= R$

The matrix A is called the factor loading matrix of the PCA.

Calculation of the proportion of variance accounted for by the first principal component C_j

Variance accounted for by $C_j = \sum_i a_{ij}^2 = \lambda_j$ (where λ_j = j^{th} eigenvalue)

The proportion of variance explained $P_j = \lambda_j / \sum_i \lambda_i$

If $P_j < k_1$ and/or $\lambda_2 > k_2$, then the group must be separated. (k_1 and k_2 are predetermined parameters.)

1.3. Orthoblique rotation

Before explaining the theory underlying this method, a brief overview of the orthogonal and oblique methods is presented.

In the orthogonal case, matrix A (factor loadings) is transformed by means of an orthogonal rotation matrix T (where T'T=TT'=I). Let B=AT, where B becomes the new orthogonal solution.

Algebraically, the correlation between a variable (zi) and an axis (Fj) is expressed as follows :

$S = A\Phi \quad (s_{ij} = \rho(z_i . F_j))$

$$\rho(z_i . F_j) = \sum_{k=1}^{v} a_{ik} \rho(F_k . F_j)$$

Where v is the number of variables.

Since the axes are perpendicular, the correlation between two axes is defined as follow :

$\sigma(F_i . F_j) = 1$ *if* $i = j$

$= 0$ *otherwise*

$\Rightarrow \Phi = I$

$\Rightarrow S = A \quad . \quad s_{ij} = a_{ij}$

In the oblique case, the matrix A is transformed by means of a rotation matrix constructed so as to obtain a solution in which the axes are not perpendicular. The matrix of the new coefficients obtained is called P to distinguish them from the old coefficients A. With an oblique rotation, the axes are correlated with each other ($F \neq I$). Then : S = PΦ

1.4. Orthoblique case

The first step is to apply the quartimax rotation method [3] to the first two eigenvectors so as to obtain the orthogonal rotation matrix T ($TT'=T'T=I$). The quartimax method maximizes a function of the 4th power of the eigenvectors in order to determine the angle of rotation .

Matrix T is then defined as follows :

$T = t_{11} = \cos(\theta)$

$t_{12} = \sin(\theta)$

$t_{21} = -\sin(\theta)$

$t_{22} = \cos(\theta)$ (where Θ is derived using the quartimax method)

$t_{ij} = 1$ *for* $i=j$ $i,j=3,\ldots,v$

$t_{ij} = 0$ *Otherwise*

To complete the rotation, a diagonal matrix that will standardize the expression $T'\Lambda T$ so as to obtain the correlation matrix with 1's on the diagonal must be found. The matrix used to perform this oblique rotation is D^{-1}.

Calculation of D^{-1}

Recall that $T' \Lambda T = [\sigma^2]$ and

that $\Phi = D^{-1} T' \Lambda T D^{-1}$ (Harman 1976)

We want $\Phi_{ij} = 1$ if i=j

Let define D^2 = Diagonal [$T' \Lambda T$] i.e. the σ^2 on the diagonal Then :

$$D_i^2 = \sigma_i^2$$

$$D_i = \sigma_i$$

$$\frac{1}{D_i} = \frac{1}{\sigma_i}$$

Implying D^{-1} = [diagonal $(T' \Lambda T)]^{-1/2}$

$$\Rightarrow \quad D^{-1}T'\Lambda T D^{-1} = \frac{\sigma_{ij}^2}{\sigma_i \sigma_j}$$

$$if \quad i=j \Rightarrow \frac{\sigma_i^2}{\sigma_i \sigma_i} = 1$$

Once T and D have been calculated, the complete oblique solution is defined as follows :

$P = Q\ T\ D$ (new coefficients)

$\Phi = D^{-1}\ T'\ \Lambda\ T\ D^{-1}$ (matrix of correlation between the axes)

$S = Q\ \Lambda\ T\ D^{-1}$ (matrix of correlation between the variables and the axes)

The Sij are used to associate each of the variables on either the first or the second axis. The results of this step gives two unconnected groups of variables. The entire procedure is applied again to the groups until they all meet the criteria set for P_1 and λ_2.

The SAS VARCLUS procedure uses a similar technique to carry out the orthoblique rotation.

APPENDIX 2

This appendix presents a sample of the COM standard long questionnaire. The variables used during the studies and their identification number are provided on the questionnaire.

Statistics Canada Statistique Canada

Industry Division

Annual Survey of Manufactures, 1987

☐ Si vous préférez remplir ce questionnaire en français, veuillez l'indiquer ici et retourner à la Division de l'industrie, Statistique Canada, Ottawa K1A 0T6 ou téléphoner (613) 951-9849

Form CM 5—3100—1.1

In all correspondence concerning this questionnaire please quote this number

Standard

Mailing Address (Please correct if necessary)

Physical Location of Establishment (Please correct if necessary)

REPORTING INSTRUCTIONS AND AUTHORITY: Completed questionnaires must be returned within 60 days of receipt. The enclosed Guide is designed to assist in the completion of this report. Instructions are numbered to correspond to the numbers on the Form. Please keep a completed copy of this Form for future reference. Collected under Authority of the Statistics Act.

INFORMATION SHARING AGREEMENTS

To reduce response burden and to ensure more uniform statistics, Statistics Canada has entered into agreements with various government departments and agencies for the sharing of data.

Under Section 10 of the Canada Statistics Act, with the provincial statistical agencies of Newfoundland, Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan, Alberta and British Columbia in respect of establishments located within the boundaries of their respective province. The Statistics Acts of these provinces include substantially the same provisions for confidentiality and penalties for disclosure of confidential information as the Canada Statistics Act.

Under Section 11 of the Canada Statistics Act, with the Prince Edward Island Department of Finance and Tourism for all establishments located within the province. Section 11 agreements shall not apply to your 1987 Survey of Manufactures report(s) if an authorized officer or person of your company objects in writing to the Chief Statistician and mails that letter to the Industry Division, of Statistics Canada together with the completed questionnaire.

CONFIDENTIALITY

Statistics Canada is prohibited by law from publishing any statistics which would divulge information obtained from this survey that relates to any identifiable business without the previous written consent of that business. The data reported on this questionnaire will be treated in strict confidence, used for statistical purposes and published in aggregate form only. The confidentiality provisions of the Statistics Act are not affected by either the Access to Information Act or any other legislation.

1.9 **REPORTING YEAR** — Report must cover your most recent financial year ending between April 1, 1987 and March 31, 1988.

from Day Month 1 9 8 Year to Day Month 1 9 8 Year

1.3.1 Did this establishment operate at any time during the reporting year as defined in 1.9 above? Yes 1 ☐ No 2 ☐

If "No" please provide a brief explanation _____

_____ and complete certification below and return this questionnaire.

Date

1.3.2 Did this establishment go out of business during the reporting year? Yes 1 ☐ _____ No 2 ☐

1.3.3 Did any change of ownership occur during the reporting year? Yes 1 ☐ _____ No 2 ☐

If "Yes" provide information for the full reporting year. If this is not possible, report for the period operated and give name and address of person to contact for the balance of the data.

Name _____ Address _____

1.6.1 Type of organization (check one)

1 ☐ Individual Ownership
2 ☐ Partnership
3 ☐ Incorporated Company
4 ☐ Co-operative

Date

1.6.2 Does the type of organization reported in 1.6.1 represent a change from your last report? Yes 1 ☐ _____ No 2 ☐

CERTIFICATION — I certify that the information contained herein is complete and correct to the best of my knowledge and belief.

Signature of authorized person	Title	Date

Name of person to contact (please print)	Telephone		
	Area code	Number	Ext.

Address including postal code (if different from mailing address above)	Postal code	Telex

STANDARD TB/CT — REG. 8100 000

Canada

1.7 **NATURE OF BUSINESS** (describe briefly)

1.7.1 _____

1.7.2 Is this a change from last year? Yes 1 ☐ No 2 ☐

1.8 **HEAD OFFICES AND ANCILLARY UNITS OF MULTI-ESTABLISHMENT FIRMS**

1.8.1 Does this establishment have a Canadian Head or Executive Office * whose operations can be reported separately? Yes 1 ☐ No 2 ☐ If "Yes" give name and address

Name _____ Address _____

1.8.5 Is this establishment served by any ancillary unit(s) * that also serve (an) other establishment(s) of your firm? Yes 1 ☐ No 2 ☐

* Data for such units should not be included in this report.

2. **INVENTORIES** at book value, including those on consignment in Canada (refer to Instruction 2 in the Reporting Guide)	Inventory for period covered by this report	
	Opening Canadian dollars (omit cents)	Closing Canadian dollars (omit cents)
2.1.1 Do these figures include inventory held but not owned? Yes 1 ☐ No 2 ☐		
Manufacturing Inventory		
2.1.2 Inventory of fuel	2.1.21	2.1.22
2.1.3 Inventory of raw materials, purchased components and supplies	2.1.31	2.1.32
2.1.5 Inventory of goods in process	2.1.51	2.1.52
2.1.6 Inventory of finished products	2.1.61	2.1.62
Non-manufacturing Inventory		
2.2 Inventory of goods purchased for resale in same condition as purchased	2.2.01	2.2.02
2.3 Other non-manufacturing inventory (specify)	2.3.01	2.3.02
2.5 Total inventory of this establishment	2.5.01	2.5.02

3 **UNFILLED ORDERS** (refer to Instruction 3 in the Reporting Guide)

3.1 Report value (or give your best estimate) as of December 31, 1967 — 3.1.0

3.2 Do you normally have a backlog (not shipping backlog) of unfilled orders? Yes 1 ☐ No 2 ☐

5 **CONSUMPTION OF PURCHASED FUEL AND ELECTRICITY** (refer to Instruction 5 in the Reporting Guide)

BASIS OF VALUATION FOR PURCHASED FUEL AND ELECTRICITY

5.0.2 Are you reporting consumption as requested? Yes 1 ☐ No 2 ☐

	Commodity code for Statistics Canada use	Cost at this establishment Canadian dollars (omit cents)
5.1 Coal	261 6	
5.2 Natural gas	263 1	
5.3 Motor gasoline (excluding aviation)	431 2	
5.4 Kerosene, stove oil (No. 1 fuel oil)	432 2	
5.5 Diesel oil	432 3	
5.6 Light fuel oil (Nos. 2 and 3)	432 4	
5.7 Heavy fuel oil (Nos. 4, 5 and 6)	432 5	
5.8 Liquefied petroleum gases (propane, butane, etc.)	436 1	
5.9 Electricity purchased (include service charge)	497 1	
5.10 Steam	497 2	
5.10.1 Other fuel purchased and used (include aviation gasoline, etc.) (Please specify)		
5.11 Total fuel and electricity		5.11.0

STANDAR

SELECTED MANUFACTURING INPUTS

6. **Raw materials, components, containers, supplies, etc., purchased and used in manufacturing operations**
(refer to Instruction 6 in the Reporting Guide) – Do not include materials, etc. produced by this establishment for its own use.

BASIS OF VALUATION FOR RAW MATERIALS, ETC.	Commodity code for Statistics Canada use	Unit of measure	Quantity used	Total cost at this establishment Canadian dollars (omit cents)
6.0.2 Are you reporting usage as requested? Yes 1 ☐ No 2 ☐				
6.1 Raw materials and components purchased and used in manufacturing operations (specify main items)				
6.2 Sub-total of items in 6.1				6.2.0

6.3 Non-returnable containers and other shipping and packaging materials purchased and used for goods of own manufacture (specify main items)	Statistics Canada use	
6.4 Sub-total of items in 6.3.		6.4.0
6.6 Total value of operating, maintenance and repair supplies purchased and used in manufacturing operations, excluding fuel		6.6.0
6.7 Total of raw materials, components, supplies, etc. (6.2 + 6.4 +6.6)		6.7.0
6.8 Amount paid out to other establishments for work done on materials owned by this establishment (refer to Instruction 6.8 in the Reporting Guide). Include amounts paid to "outside pieceworkers" in the clothing industries		6.8.0
6.9 Total of raw materials, containers, supplies and amount paid for work done (6.7 + 6.8)		6.9.0

SELECTED NON-MANUFACTURING INPUTS	Total cost at this establishment Canadian dollars (omit cents)
7. **Merchandising and construction activities, etc.** (refer to Instruction 7 in the Reporting Guide)	
7.1 **Purchases of goods from other establishments for resale in same condition as purchased** (include transfers of such goods from other establishments of your company) (report sales of such goods in question 9.1)	7.1.0
7.2 **Purchased materials and supplies used in new construction produced by own labour force for own use** (only those items charged to Fixed Assets Accounts which are reported in question 9.2)	7.2.0
7.3 **Purchased materials and supplies used in production of any machinery and equipment for own use by own labour force** (only those items charged to Fixed Assets Accounts which are reported in question 9.3)	7.3.0
7.4 Office supplies purchased and used	7.4.0
7.5 All other purchased materials and supplies used by this establishment	7.5.0
7.6 Sub-total of items in 7	7.6.0
7.7 Grand total of selected manufacturing and non-manufacturing inputs (6.9 + 7.6)	7.7.0

SELECTED MANUFACTURING OUTPUTS

8. Shipments of goods of own manufacture (refer to Instruction 8 in the Reporting Guide)

BASIS OF VALUATION OF SHIPMENTS	Commodity code for Statistics Canada use	Unit of measure	Quantity shipped	Net value of shipments, excluding sales taxes, excise duties and excise taxes, shipping charges by common or contract carriers and net of any sales discounts, allowances, etc. Canadian dollars (omit cents)
8.0.2 Are you reporting shipments as requested? Yes 1 ☐ No 2 ☐				
8.0.3 Are reported Values at Final Selling Price? Yes 1 ☐ No 2 ☐ If "No" please specify basis of valuation:				
8.0.4 _____ ☐				
8.1 Products shipped (specify main items)				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
.....				
8.2 Total value of shipments of goods of own manufacture (total of items in 8.1)				8.2.0

8. MANUFACTURING OUTPUTS – concluded

	Net value of shipments excluding sales taxes, excise duties and excise taxes, shipping charges by common or contract carriers and net of any sales discounts, allowances, etc. Canadian dollars (omit cents)
8.3 Less adjustments for the following items if you were not able to exclude them from the value of the individual products in section 8.1	
8.3.1 Total payments for shipping charges by common or contract carriers	8.3.1
8.3.2 Total payments of sales taxes, excise duties and excise taxes	8.3.2
8.3.3 Total amounts of discounts, sales allowances and returned sales	8.3.3
8.4 Total adjustments (sum of items in 8.3)	8.4.0
If the amounts reported above include any incurred in connection with goods purchased for resale (see 9.1) please check here ☐	
8.5 Adjusted value of shipments of goods of own manufacture (8.2 less 8.4 or 8.2 if 8.4 is zero)	8.5.0
8.6 Amount received in payment for work done on materials and products owned by other establishments (including those from any other establishment of your own company)	8.6.0
8.7 **Total value of shipments of goods of own manufacture and amount received for work done (8.5 + 8.6)**	8.7.0

9. SELECTED NON-MANUFACTURING OUTPUTS (refer to Instruction 9 in the Reporting Guide)

		Canadian dollars (omit cents)
→	9.1 **Total value of shipments of goods purchased and sold in the same condition as purchased** (purchases of such goods should be reported in question 7.1)	9.1.0
	Specify below the major products included in this value and give estimate of percentage which each represents of this value (9.1).	
	Name of product ____ Estimated %	
	____ ____	
	____ ____	
	____ ____	
→	9.2 **Book value of new construction by own labour force for own use** (only that amount charged to the Fixed Assets Accounts — this should include at least material costs reported in question 7.2 and labour included in 14.1.4)	9.2.0
→	9.3 **Book value of machinery and equipment manufactured by own labour force for own use** (only that amount charged to the Fixed Assets Accounts — this should include at least material costs reported in question 7.3 and labour included in 14.1.4)	9.3.0
→	9.5 **Revenue from lease or rental of machinery and equipment manufactured by this establishment**	9.5.0
→	9.6 **All other revenue from products and services** (exclude non-operating revenues such as interest, dividends and other rental income, etc.)	9.6.0
→	9.7 **Total of items in 9**	9.7.0
→	10. **Grand total of manufacturing and selected non-manufacturing outputs** (8.7 + 9.7)	10.0.0
	SUPPLEMENTARY	
→	11. **Revenue from lease or rental of property** (lands, buildings, offices, etc.)	11.0.0
→	12. **Revenue from lease or rental of machinery and equipment other than that included in 9.5 above** (i.e. from machinery of all kinds, engines, trucks of all types, trailers, tractors, other equipment, etc.)	12.0.0

13. **EXPORTS**

Please estimate the percentage of your Total Value of Outputs (Line 10, above) which are shipped:

13.61 within Canada ____ %

13.62 to the U.S.A. ____ %

13.63 to other countries ____ %

Total 100%

APPENDIX 3

Study of correlation coefficients

The grouping method uses a correlation matrix as input. Studies have been carried out to evaluate the behaviour of three correlation coefficients : those of Pearson, Spearman and Kendall.

Pearson correlation's coefficient =

$$\frac{\Sigma_i \ (x_i-x) \ (y_i-y)}{\sqrt{(\Sigma_i \ (x_i-x)^2 \ \Sigma_i \ (y_i-y)^2)}}$$

Spearman correlation's coefficient =
(non-parametric coefficient based on ranks)

$$\frac{\Sigma_i \ (r_i-r) \ (s_i-s)}{\sqrt{(\Sigma_i \ (r_i-r)^2 \ \Sigma_i \ (s_i-s)^2)}}$$

where r_i is the rank of the i^{th} value of x
and s_i is the rank of the i^{th} value of y

Kendall correlation's coefficient =
(non-parametric coefficient based on signs)

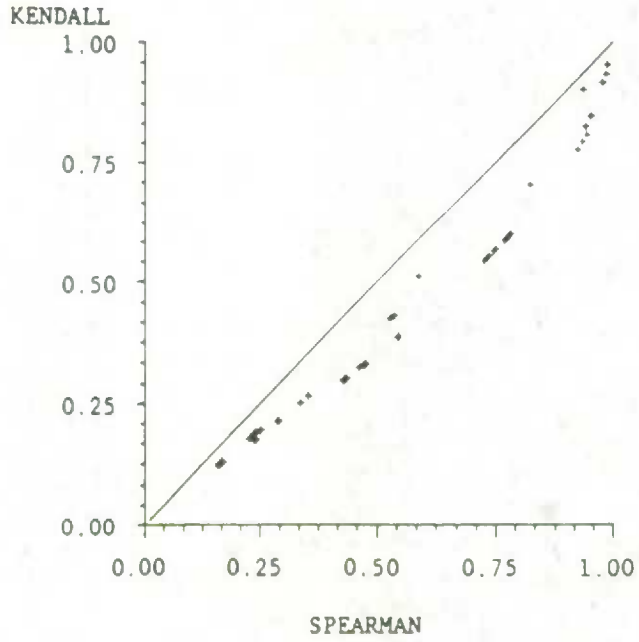
$$\frac{S}{n(n-1)\div 2}$$ where n = number of observations

1. Arrange observations in increasing order of variable x
2. Compare each y with all y located below. A pair of y (a y compared to a y located below it) is in natural order if the y situated lower is greater than the other. If not, a pair is in inverse natural order.
3. P is the number of pairs in natural order, Q is the number of pairs in inverse natural order.
4. S = P-Q

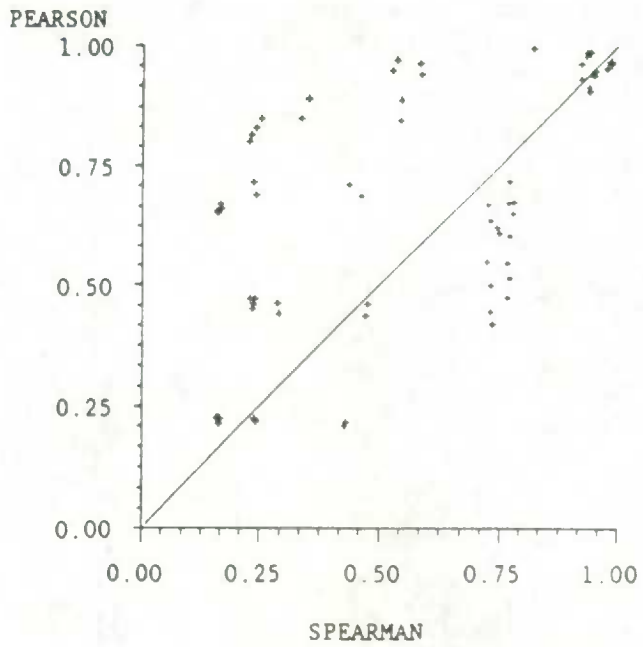
   n(n-1)÷2 = total number of pairs

The correlation matrix was calculated for the same set of data for each of the three coefficients. In order to compare their behaviour, graphs representing the values of one correlation coefficient as a function of the other have been made. A line of slope one was added to the graphs to help for their interpretation.

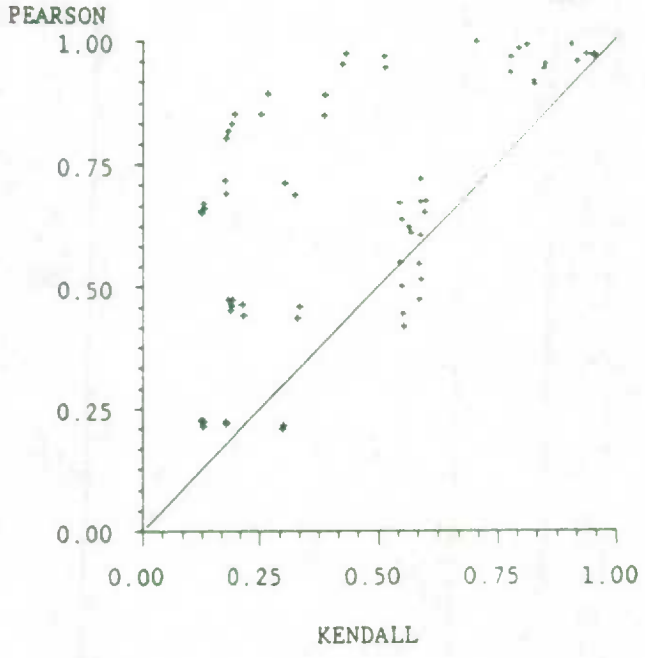
KENDALL v SPEARMAN



PEARSON v SPEARMAN



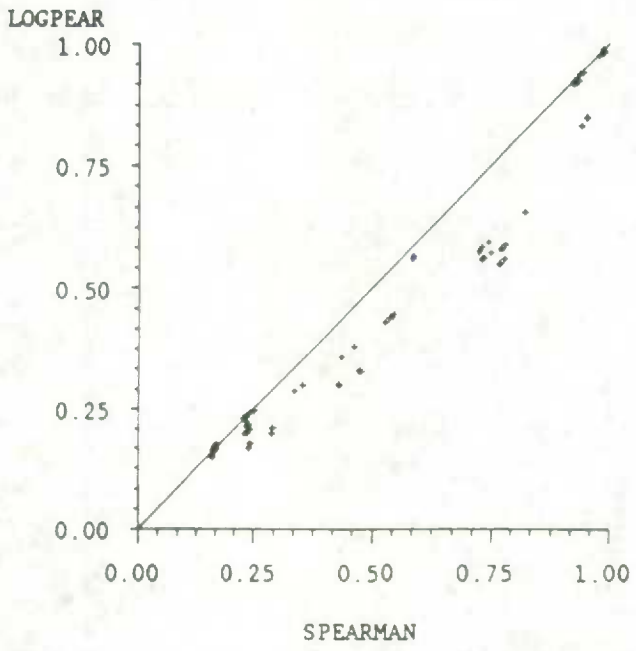
PEARSON v KENDALL



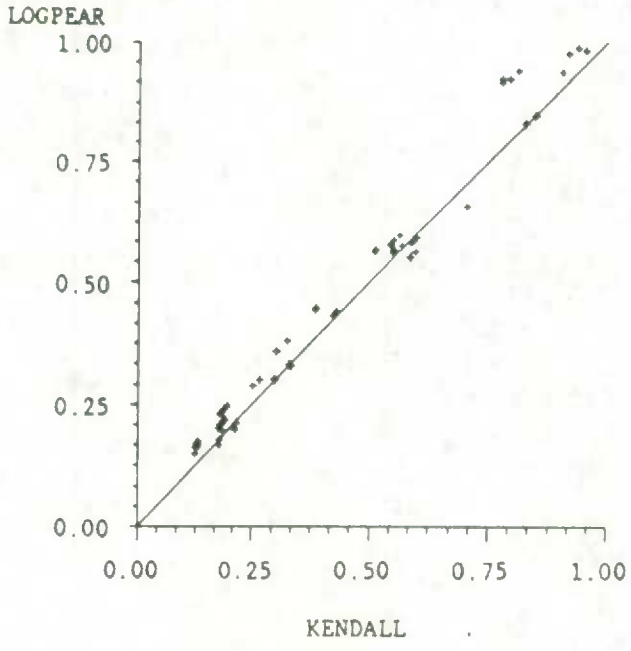
Analysing the first graph (Kendall vs Spearman), it can be observed that the two coefficients behave similarly. Kendall seems, however, to be slightly stricter (<) than Spearman. The second and third graphs (Pearson vs Spearman and Pearson vs Kendall) are similar. Pearson tends in general to be larger than the other two coefficients.

As mentionned in section 1.3.2, a data transformation is needed to make Pearson less dependent on large values. Graphs two and three have been redrawn, this time using Pearson's correlation calculated from the logarithm of the variables. Here are the results :

PEARSON (LOG) vs SPEARMAN



PEARSON (LOG) v KENDALL



By transforming the data by means of the logarithm, the impact of large values has been reduced, thereby making Pearson's correlation coefficient stricter. Its behaviour is similar to that of the two non-parametric coefficients.

APPENDIX 4

List of the groups identified during the empirical study on the grouping method.

Results obtained for the 1985 COM (30 groups). The letter C stands for 1985 COM (for the variable number, refer to appendix 2).

Group 1

C6.9.0
C7.7.0
C8.2.0
C8.5.0
C8.7.0
C10.0.0

Group 2

C14.2.11
C14.2.12
C14.2.13
C14.2.14
C14.2.15
C14.3.04
C14.3.05

Group 3

C5.11.0
C6.2.0
C6.7.0
C14.1.21
C14.1.22
C14.1.24
C14.1.25
C14.1.91
C14.1.92

Group 4

C14.2.21
C14.2.22
C14.2.23
C14.3.01
C14.3.02
C14.3.03

Group 5

C2.1.31
C2.1.32
C2.5.01
C2.5.02

Group 6

C7.1.0
C7.6.0
C9.1.0
C9.7.0

Group 7

C8.3.1
C8.3.3
C8.4.0

Group 8

C14.1.11
C14.1.12
C14.1.13

Group 9

C2.1.21
C2.1.22

Group 10

C2.1.51
C2.1.52

Group 11

C2.1.61
C2.1.62

Group 12

C2.2.01
C2.2.02

Group 13

C2.3.01
C2.3.02

Group 14

C7.2.0
C9.2.0

Group 15

C7.3.0
C9.3.0

Group 16

C14.1.23
C14.1.93

Group 17

C14.1.41
C14.1.42

Group 18

C3.1.0

Group 19

C6.4.0

Group 20

C6.6.0

Group 21

C6.8.0

Group 22

C7.4.0

Group 23

C7.5.0

Group 24

C8.3.2

Group 25

C8.6.0

Group 26

C9.5.0

Group 27

C9.6.0

Group 28

C11.0.0

Group 29

C12.0.0

Group 30

C14.1.43

Results obtained for the 1984 COM (28 groups). The letter Q stands for the 1984 COM.

24 of the 28 groups identified for the 1984 COM are exactly the same then those identified for the 1985 COM. The four groups that have changed are now presented.

Group 1

Combination of group 1 and group 20 of the 1985 COM

Q6.6.0
Q6.9.0
Q7.7.0
Q8.7.0
Q10.0.0

Group 2

Combination of group 2 and group 4 of 1985 COM

Q14.2.11
Q14.2.12
Q14.2.13
Q14.2.14
Q14.2.15
Q14.2.21
Q14.2.22 Note : Q14.3.04 and Q14.3.05 did not exist in the 1984 COM.

Group 3

Combination of group 6 and group 12 of 1985 COM

Q2.2.01
Q2.2.02
Q7.1.0
Q7.6.0
Q9.1.0
Q9.7.0

Group 4

Two variables from group 1 of 1985 COM.

Q8.2.0
Q8.5.0

Results obtained for the grouping of the 146 variables comming from both the 1984 and 1985 COM. The letter C stands for the 1985 COM while the letter Q stands for the 1984 COM.

Group 1

C6.9.0
C7.7.0
C8.7.0
C10.0.0
Q6.9.0
Q7.7.0
Q8.7.0
Q10.0.0

Group 2

C8.2.0
C8.5.0
Q8.2.0
Q8.5.0

Group 3

C14.2.11
C14.2.12
C14.2.14
C14.2.15
C14.3.04
C14.3.05
Q14.2.11
Q14.2.12
Q14.2.14
Q14.2.15
Q14.3.01
Q14.3.02

Group 4

C5.11.0
C14.1.21
C14.1.22
C14.1.24
C14.1.25
C14.1.91
C14.1.92
Q5.11.0
Q14.1.21
Q14.1.22
Q14.1.24
Q14.1.25
Q14.1.91
Q14.1.92

Group 5

C6.2.0
C6.7.0
Q6.2.0
Q6.7.0

Group 6

C14.2.21
C14.2.22
C14.2.23
C14.3.01
C14.3.02
C14.3.03

Group 7

Q14.2.21
Q14.2.22
Q14.2.23

Group 8

C2.1.31
C2.1.32
Q2.1.31
Q2.1.32

Group 9

C2.5.01
C2.5.02
Q2.5.01
Q2.5.02

Group 10

C7.1.0
C7.6.0
C9.1.0
C9.7.0
Q7.1.0
Q7.6.0
Q9.1.0
Q9.7.0

Group 11

C8.3.1
C8.4.0
Q8.3.1
Q8.4.0

Group 12

C8.3.3
Q8.3.3

Group 13

C14.1.11
C14.1.12
C14.1.13
Q14.1.11
Q14.1.12
Q14.1.13

Group 14

C2.1.21
C2.1.22
Q2.1.21
Q2.1.22

Group 15

C2.1.51
C2.1.52
Q2.1.51
Q2.1.52

Group 16

C2.1.61
C2.1.62
Q2.1.61
Q2.1.62

Group 17

C2.2.01
C2.2.02
Q2.2.01
Q2.2.02

Group 18

C2.3.01
C2.3.02
Q2.3.01
Q2.3.02

Group 19

C7.2.0
C9.2.0
Q7.2.0
Q9.2.0

Group 20

C7.3.0
C9.3.0
Q7.3.0
Q9.3.0

Group 21

C14.1.23
C14.1.93
Q14.1.23
Q14.1.93

Group 22

C14.1.41
C14.1.42
Q14.1.41
Q14.1.42

Group 23

C3.1.0
Q3.1.0

Group 24

C6.4.0
Q6.4.0

Group 25

C6.6.0
Q6.6.0

Group 26

C6.8.0
Q6.8.0

Group 27

C7.4.0
Q7.4.0

Group 28

C7.5.0
Q7.5.0

Group 29

C8.3.2
Q8.3.2

Group 30

C8.6.0
Q8.6.0

Group 31

C9.5.0
Q9.5.0

Group 32

C9.6.0
Q9.6.0

Group 33

C12.0.0
Q12.0.0

Group 34

C14.1.43
Q14.1.43

Group 35

C11.0.0
Q14.1.93

Group 36

C14.2.13
Q14.2.13
Q14.3.03

APPENDIX 5

Analysis of the linear regression model adjusted for 1985 Census Of Manufactures data (Québec)

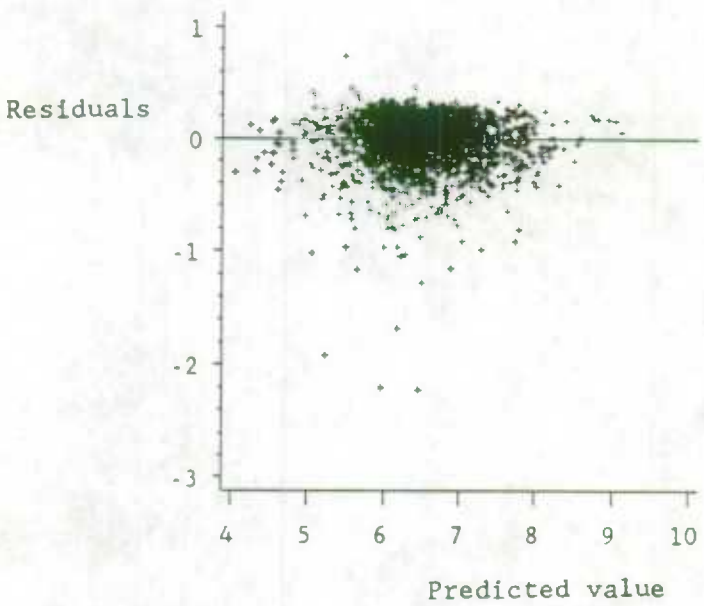
Model :

Log(inputs) = -.54 + 1.03 log(outputs) + e

R^2 = 0.88
F = 21213.79

There are several ways of verifying the homoscedasticity (equality of variances) of the residuals in a regression model. One of the simplest is to plot the residuals as a function of the predicted values and study its shape.

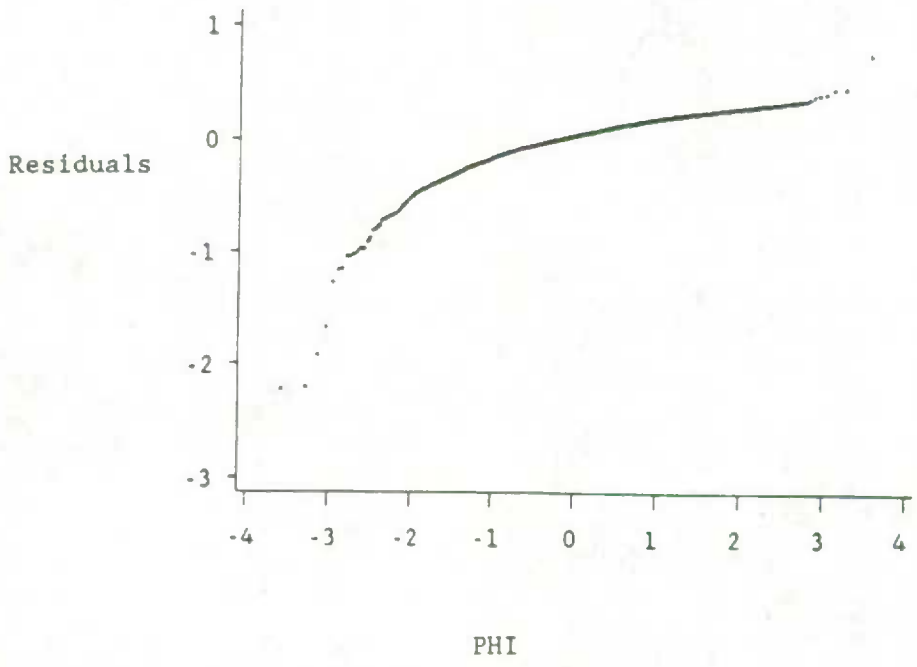
GRAPH OF RESIDUALS AS A FUNCTION OF PREDICTED VALUES RESIDUALS



From the results obtained, the residual variances can be assumed to be equal. The dots on the graph are randomly distributed around a line of slope zero, and do not present any systematic pattern. There does, however, seem to be an upper bound to the residuals. This is due once again to the natural constraint existing between manufacturing inputs and outputs.

For a model to be adequate, the residuals must be distributed according to the Normal distribution. Henri's line, or a normal probability plot, was traced to verify this hypothesis.

RESIDUALS AS HENRI'S LINE RESIDUALS



The graph takes the form of a line except at the two extremes. The normality of the residuals can thus be assumed. The model adjusted for manufacturing inputs and outputs variables is thus adequate and can be used to detect questionable units.

WORKING PAPER NO. BSMD-90-005E/F

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-90-005E/F

DIRECTION DE LA MÉTHODOLOGIE

ANALYSES SUR LA CLASSIFICATION DES VARIABLES ET SUR LA DÉTECTION DES DONNÉES SUSPECTES

par

France Bilocq et J.-M. Berthelot

Mars 1990

ANALYSES SUR LA CLASSIFICATION DES VARIABLES ET SUR LA DÉTECTION DES DONNÉES SUSPECTES

par

France Bilocq et J.-M. Berthelot

Mars 1990

TABLE DES MATIÈRES

APPENDICES

0. INTRODUCTION

La vérification dans la fonction générale de collecte et de saisie des données n'est pas une étape isolée. Elle fait partie d'un processus intégré qui se compose des trois points suivants:

1) Dans un premier temps, des groupements naturels de variables sont identifiés, permettant d'optimiser l'application des procédures de vérification.

2) Par la suite les dossiers suspects sont identifiés à l'aide de techniques statistiques. Un dossier suspect est un dossier ayant une ou plusieurs variables dont les valeurs dévient significativement du reste de la population.

3) Finalement, les dossiers suspects sont classifiés (fonction de caractérisation) pour fins de confirmation et/ou correction via les procédures de suivis ou pour imputation via le système généralisé d'imputation.

Le but de ce document est de présenter les résultats des études empiriques réalisées, concernant le groupement des variables et la détection des données suspectes. La fonction de caratérisation fait présentement l'objet de recherche plus approfondies. Des résultats seront présentés ultérieurement.

1. MÉTHODE DE CLASSIFICATION DES VARIABLES

Afin de faciliter le développement de la méthodologie pour la vérification des données, on fait appel aux techniques d'analyse de données pour concevoir le schéma de vérification pour les variables quantitatives des enquêtes économiques. L'analyse factorielle est une branche de l'analyse des données qui se concentre sur l'étude des relations internes d'un ensemble de variables. La méthode de classification des variables décrite ci-dessous fait appel à cette théorie.

La méthode de classification est un outil statistique qui sert à identifier des groupements naturels de variables. Le but de la méthode est de partitionner un ensemble de variables en deux ou plusieurs groupes disjoints. Chacun des groupes formés contient des variables qui sont fortement corrélées entre elles. Par la suite, un patron de vérification faisant intervenir uniquement les variables à l'intérieur d'un groupe est développé afin de minimiser le nombre de vérifications croisées (entre deux variables).

1.1 DESCRIPTION DE LA MÉTHODE DE CLASSIFICATION

La méthode de classificaton considérée partitionne un ensemble de variables en groupes disjoints en se basant sur leur matrice de corrélation. Les groupes sont formés de façon à maximiser la variance expliquée par la première composante principale de chaque groupe tout en s'assurant que la variance expliquée par la deuxième composante principale n'est pas trop élevée. On contrôle ces deux contraintes à l'aide de paramètres.

Étapes de la méthode de classification :

Au départ, la méthode considère toutes les variables comme faisant partie du même groupe. Par la suite les étapes suivantes sont répétées jusqu'à ce que toutes les conditions soient remplies.

1) La première étape consiste à effectuer, de façon indépendante l'analyse en composantes principales de la matrice de corrélation de chacun des groupes de variables.

2) Un groupe donné est subdivisé si la proportion de variance expliquée par sa première composante principale est insuffisante ou si la proportion de variance expliquée par sa deuxième composante principale est trop grande. Autrement, un groupe répondant aux critères n'est pas séparé.

3) Un groupe est séparé en effectuant une variante de la rotation orthoblique proposée par H.H. Harman (1976) [2] et W. Harris et F. Keiser (1964) [3]. Cette technique se nomme ainsi puisqu'on obtient un solution oblique à partir d'une rotation orthogonale et d'une matrice définie positive. Une solution oblique implique que les axes de référence ne sont pas orthogonaux.

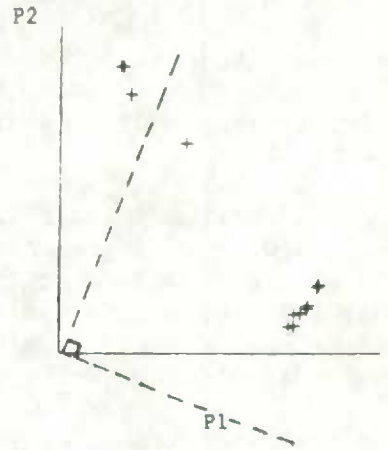
   Un des principes de base de l'analyse factorielle est de toujours être à la recherce d'une solution simple. Par solution simple on entend une solution pour laquelle chaque variable est bien représentée par peu d'axes (de préférence un seul). Or, une telle solution ne se trouve pas toujours automatiquement par l'analyse factorielle d'une matrice de corrélation. C'est pourquoi on fait appel à la rotation.

   L'analyse en composantes principales fournit une des solutions possibles d'axes de référence parmi une infinité. La rotation consiste à faire pivoter ces axes de façon à obtenir une meilleure représentation des variables sur ceux-ci. Cette étape permet de trouver une solution plus claire qui explique plus adéquatement la variabilité d'un groupe de variables.

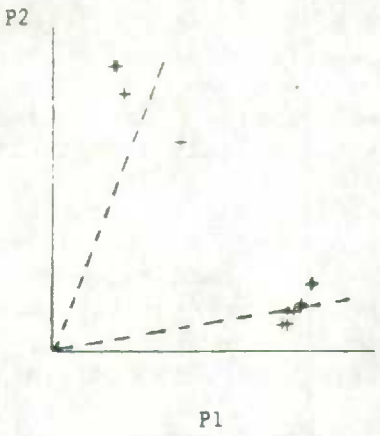
   Pourquoi rechercher une solution oblique plutôt qu'orthogonale ? En imposant un cadre de référence orthogonal on impose aussi la même contrainte aux axes qui le compose. Par exemple, on sait qu'il existe un lien entre les variables concernant les produits manufacturiers et les variables concernant les produits non-manufacturiers (Recensement des manufacturiers). Cependant, ce lien n'est pas assez fort pour justifier le regroupement de ces deux catégories de variables de façon à ce que l'ensemble des variables soient bien expliquées par un seul axe.

Si on impose une rotation dont la solution est orthogonale, on obtient que l'un ou l'autre des groupes est bien expliqué par un des axes au détriment de l'autre. En permettant l'obtention d'une solution oblique on peut alors bien expliquer les deux groupes de variables simultanément à l'aide d'axes différents. Afin de mieux visualiser la situation transportons-nous dans le graphique des variables représentées dans le plan des deux premières composantes principales.

GRAPHIQUE solution orthogonale



GRAPHIQUE solution OBLIQUE



Note : Les données utilisées pour ces graphiques proviennent du Recensement des Manufacturiers de 1985.

La solution oblique est plus simple à interpréter. Opter pour une telle solution permet d'obtenir une classification des variables plus détaillée.

Le nombre de groupes obtenu par la méthode de classification se contrôle à l'aide des paramètres spécifiés quant à la proportion de variance expliquée par les deux premières composantes principales. En effet, plus le paramètre contrôlant la première composante principale est grand ou plus celui contrôlant la deuxième composante principale est petit, alors plus le nombre de groupes formés est grand.

Finalement, pour séparer un groupe en deux, on regarde les corrélations entre les variables et les deux premiers axes obtenus par la rotation orthoblique. Chaque variable est jumelée avec l'axe avec lequel elle est le plus corrélée pour ainsi obtenir deux groupes. Étant donnée que la rotation orthoblique fournit une solution dont les axes ne sont pas orthogonaux, la méthode de classification fournit les coefficients de corrélation entre les groupes.

4) Par la suite, un processus itératif réassigne (au besoin) les variables dans les groupes afin de maximiser la variance expliquée par la première composante de chacun des groupes.

5) Lorsque tous les groupes satisfont les critères spécifiés quant à la proportion de variance expliquée par les deux premières composantes principales alors le processus est terminé. Sinon, retour à l'étape 1.

Les détails algébriques sont présentés à l'appendice 1.

1.2 COEFFICIENTS DE CORRÉLATION

La méthode de classification des variables utilise en entrée une matrice de corrélation. Différents types de coefficients peuvent être utilisés.

Le coefficient de corrélation de Pearson mesure le degré d'association linéaire entre deux variables. Il est facile à interpréter et aussi très rapide à calculer. Cependant il est influencé par les grandes valeurs. Dans les enquêtes économiques, cette caratéristique tend à entrainer une surestimation de la corrélation entre les variables.

Les coefficients de corrélation de Spearman et de Kendall sont des coefficients non-paramétriques. Ils peuvent mesurer un degré d'association qui n'est pas nécessairement linéaire. Par exemple une fonction monotone croissante entre deux variables produira des coefficients de Spearman et de Kendall d'une valeur de un. Ces deux coefficients sont plus difficiles à interpréter et sont plus longs et plus coûteux à calculer.

Comme de façon générale, on retrouve des liens linéaires entre les variables d'une enquête économique et comme le coefficient de Pearson est mieux connu, plus facile à interpréter et plus rapide à calculer, il a été retenu pour les besoins des études empiriques. Cependant, on verra plus loin que des tranformations sont nécessaires afin de minimiser l'impact des grandes valeurs sur le coefficient de Pearson.

1.3 ÉTUDES EMPIRIQUES

1.3.1 Données utilisées

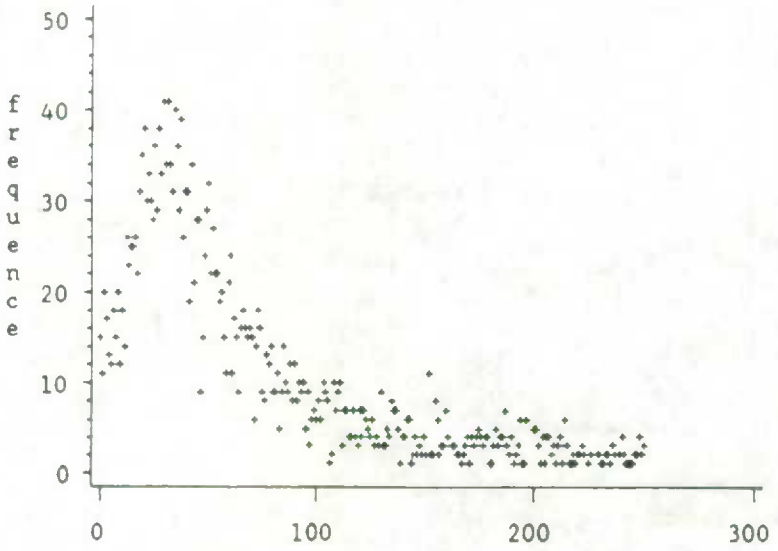
De études empiriques ont été effectuées dans le but d'étudier le comportement de la méthode de classification sur des données réelles. Lors de ces études, la procédure VARCLUS de SAS [6] été utilisée pour effectuer la classification. Cette procédure utilise une technique similaire à celle décrite précédemment.

Les données finales des questionnaires longs des recensements des manufacturiers (RM) de 1984 et 1985 ont été utilisées à cette fin. Les données finales sont celles qui ont été vérifiées, corrigées, imputées, c'est-à-dire les données utilisées lors des publications. Il existe plusieurs versions du questionnaire long du recensement des manufacturiers. Le questionnaire varie avec la classification industrielle. Seules les variables financières standard ont été analysées. Ces variables étant celles que l'on retrouve systématiquement sur chaque version du questionnaire long. Un exemplaire du questionnaire standard se retrouve à l'appendice 2.

1.3.2 Analyses préliminaires

Avant de foncer tête première dans l'application de la méthode de classification, l'analyse de la distribution de fréquence des variables a été effectuée. Le graphique suivant représente la distribution de fréquence, au niveau Canada, d'une des variables financière à l'étude.

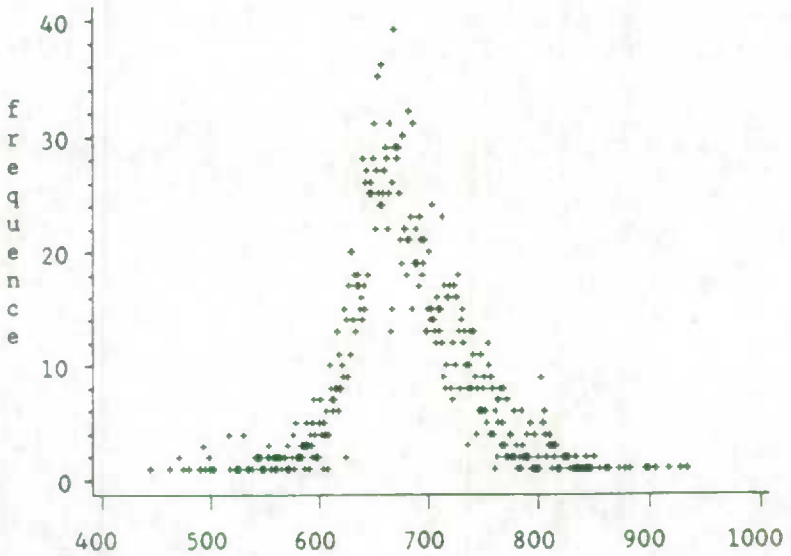
DISTRIBUTION DE FRÉQUENCE (entrées manufacturières)
(niveau Canada)



On observe que la variable a une distribution de fréquence asymétrique qui s'approche beaucoup de celle d'une loi lognormale. Il y a beaucoup de dossiers ayant de faibles valeurs et très peu ayant de grandes valeurs. Toutes les variables financières à l'étude ont une distribution de fréquence similaire. Des analyses faites à des niveaux inférieurs (provinces, classification industrielle) ont donné des résultats similaires.

La méthode de classification des variables utilise les coefficients de corrélation de Pearson. On sait que ce dernier est influencé par les grandes valeurs. On a donc transformé les données afin d'obtenir une distribution plus symétrique et ainsi atténuer l'impact des grandes valeurs sur le coefficient de corrélation. La transformation utilisée est celle du logarithme. Le graphique suivant représente la distribution de fréquence de la variable transformée.

DISTRIBUTION DE FRÉQUENCE LOG(entrées manufacturieres)
(niveau Canada)



[Log (entrées manufacturières)] (x 10^2)

La distribution est maintenant plus symétrique. Lors de l'application de la méthode de classification, le coefficient de corrélation de Pearson sera calculé à partir du Log des variables. Ainsi les corrélations obtenues seront moins influencées par les grandes valeurs et donc un peu plus sévères. Les résultats d'études plus approfondies sur les coefficients de corrélation sont présentés à l'appendice 3.

1.3.3 Résultats (méthode de classification)

La méthode de classification des variables a été appliquée sur les données du recensement des manufacturiers (RM) de 1984 et 1985. Deux études ont été effectuées.

Première étude

La première étude a été faite au niveau Canada pour les RM de 1985 et 1984 et consistait à classifier les 73 variables financières du questionnaire long, sans tenir compte de la structure du questionnaire.

Les résultats obtenus pour le RM de 1985 sont les suivants :

- Trente groupes ont été formés. La taille des groupes varie d'une à neuf variables.

- Les groupes sont cohérents c'est-à-dire que les variables présentes dans un groupe ont toutes un lien logique entre elles.

- Les résultats obtenus concordent avec les connaissances que les responsables du RM possèdent sur le comportement des variables.

Les résultats obtenus pour le RM de 1984 sont les suivants :

- Vingt-huit groupes ont été formés. Deux de moins que pour le recensement de 1985.

- L'analyse des résultats a révélé que 24 des 28 groupes formés pour le RM de 1984 sont exactement les mêmes que ceux formés pour le RM de 1985. Les quatres autres groupes proviennent de la séparation ou de la fusion des groupes formés pour le RM de 1985.

  À partir des résultats obtenus ont a supposé que les données historiques (RM 1984) et les données courantes (RM 1985) avaient un comportement similaire.

Partant de cette hypothèse, la méthode de classification a été appliquée simultanément aux 146 variables des deux recensements. Les résultats sont concluants. Trente-six groupes ont été formés. Un détail important à mentionner, les groupes sont en majorité les mêmes sauf que leur taille a doublée. Les mêmes variables des deux recensements ont été jumelées, ce qui somme toute confirme la similarité des données d'un cycle à l'autre du recensement des manufacturiers.

La liste des groupes formés lors des trois expériences précédentes est fournie à l'appendice 4.

Deuxième étude

Le but de la seconde étude était d'évaluer le comportement de la méthode lorsqu'appliquée à différents niveaux tels que Canada, provinces, et classification industrielle. Cette fois les variables à l'étude ont été choisies en fonction de la structure du questionnaire. Les douze variables les plus importantes ont été retenues. Ces variables se divisent en quatre grandes classes:

1. Stocks d'ouverture
2. Stocks de fermeture
3. Les entrées
4. Les sorties

qui à leur tour sont subdivisées en trois composantes :

a. produits manufacturiers
b. produits non-manufacturiers
c. Total des produits manufacturiers et non manufacturiers.

Les variables 1a 1b 2a et 2b ont été créées en sommant les composantes appropriées à partir du questionnaire original.

Dans un premier temps, les variables ont été classifiées au niveau Canada. Les résultats obtenus sont présentés dans le tableau suivant. Le numéro dans chaque case correspond au groupe dans lequel la variable a été classifiée.

RÉSULTATS NIVEAU CANADA

VARIABLES	Stocks ouv.	Stocks ferm.	Entrées	Sorties
Produits manufacturiers	1	1	2	2
Produits non-manufacturiers	3	3	4	4
Total	1	1	2	2

Quatre groupes ont été formés. Si on examine les résultats obtenus on se rend compte qu'ils sont cohérents. On sait pertinemment que les stocks d'ouverture et de fermeture sont fortement reliés. On retrouve un lien semblable entre les entrées et les sorties. Il est également connu que la composante manufacturière est celle qui contribue le plus au total. Les groupes sont donc représentatifs des caractéristiques spécifiques observées sur les données.

Par la suite, la méthode de classification a été appliquée par province. Pour chacunes d'entre elle (10), l'ensemble des douze variables a été séparé en quatre. On a obtenu exactement les mêmes groupes que lors de la classification au niveau Canada.

Finalement, la méthode a été appliquée au niveau de la classification industrielle à deux chiffres (SIC2). Douze SIC2 ont fait l'objet d'une analyse. Pour six d'entre eux, les résultats sont les mêmes que précédemment. Cependant pour les six autres SIC2, trois groupes ont été formés au lieu de quatre. En analysant d'un peu plus près la situation, il a été noté que le troisième groupe correspond, dans les six cas, à une combinaison de deux des quatre groupes de base. Soit une combaison des groupes un et deux ou encore une combinaison des groupes trois et quatre.

Ces résultats respectent les caractéristiques énoncées ci-haut à propos des composantes manufacturières et non-manufacturières.

NOTE : En modifiant la valeur des paramètres (variance expliquée par les deux premières composantes principales) on a obtenu, pour les 12 SIC2 à l'étude, les quatre groupes de base.

1.4 CONCLUSION

Les résultats des études empiriques sont concluants. La méthode de classification extrait facilement le patron de corrélation d'un ensemble de variables. Les variables regroupées ont toutes un lien logique entre elles et les résultats concordent avec les connaissances des responsables du recensement des manufacturiers.

Étant donné que la procédure existe déjà dans le progiciel SAS c'est une méthode qui est facile à mettre en place. De plus elle est facile à utiliser et à interpréter. C'est aussi une méthode économique puisque la transformation des données (log) et le calcul des coefficients de corrélation de Pearson sont des opérations rapide d'exécution. L'analyse multivariée s'effectue sur la matrice de corrélation. Ceci entraine aussi une économie puisque l'on travaille avec un ensemble de données réduit.

Il est important de retenir que la méthode de classification a pour but de regrouper les variables qui sont reliées entre elles pour les besoins de la vérification. Une fois que les groupes sont formés, on peut restreindre les vérifications croisées aux relations qui existent entre les variables d'un groupe tout en assurant la cohérence des variables d'un dossier. Des vérifications impliquant des variables de différents groupes seront effectuées uniquement si elles sont nécessaires pour assurer la cohérence d'un dossier. La méthode de classification fournit les coefficients de corrélation entre les groupes. Cette information pourrait dans ce cas servir de guide dans le choix des vérifications inter-groupes.

La méthode de classification des variables est un outil de travail permettant de combiner à la fois théorie statistique et expérience des données. Un des objectifs de la théorie statistique est de fournir une loi scientifique ou un modèle mathématique pour expliquer le comportement des données. Si l'on possède une connaissance préalable du comportement des variables étudiées, ces notions permettront alors de vérifier la cohérence des groupes formés. De plus, certains liens inconnus entre des variables peuvent être dévoilés grâce à l'application de cette méthode. Cependant si on ne possède pas cette connaisance ou si l'on veut s'appuyer sur des résultats d'expérience, alors la méthode de classification permet de grouper les variables de façon empirique et objective. Cette méthode permet donc d'enchâsser le processus de vérification des données dans un cadre théorique sans toutefois en oublier le caractère intuitif.

2. MÉTHODES DE DÉTECTION

Le but de cette section est de présenter les résultats des études empiriques effectuées dans le but d'évaluer le comportement de deux méthodes de détection bivariée soit la régression linéaire simple et la méthode des tendances. Les méthodes de détection multivariées feront l'objet de recherches ultérieures.

2.1 DÉFINITION

Une méthode de détection bivariée modèlise les liens existant entre deux variables soit d'un même cycle ou de cycles différents pour permettre l'identification de liens incohérents entre ces deux variables lors d'un cycle courant.

Pour évaluer si une méthode de détection est efficace ou non, les critères d'évaluation suivants ont été utilisés :

- Pour aider à mieux visualiser le comportement d'une méthode de détection bivariée, on trace le graphique représentant le nuage de point d'une variable en fontion de l'autre. L'intervalle d'acceptation obtenue par la méthode est ensuite ajoutée sur le graphique. On s'attend à ce que les bornes de rejet épousent autant que possible la forme du nuage de points.

- Les dossiers avec de grandes valeurs sont très importants. Ils ont parfois un comportement différent des autres. Une méthode de détection devrait modéliser adéquatement le comportement de ces dossiers.

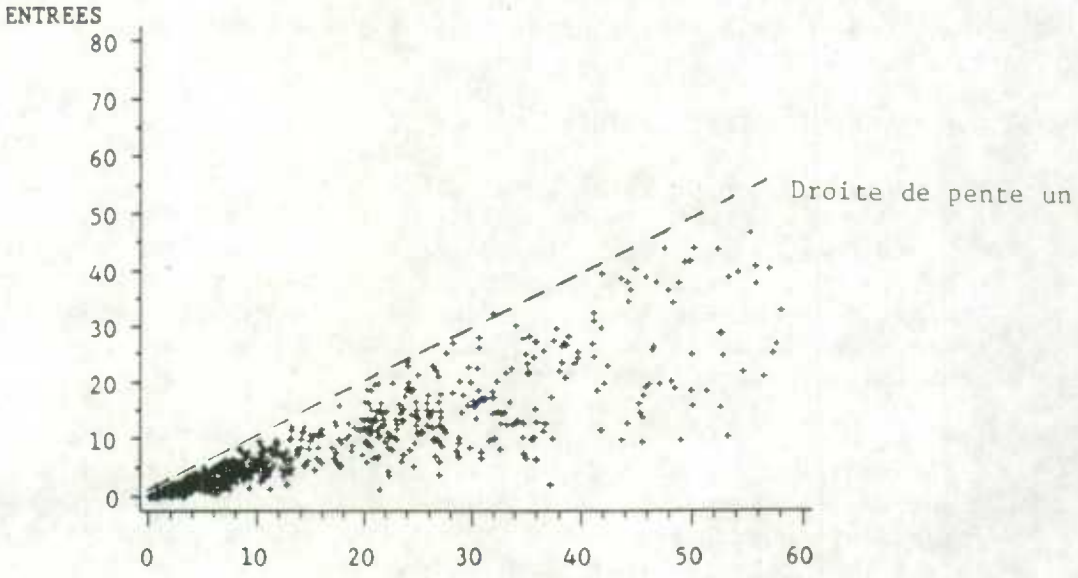
- Les variables ne sont pas toujours rapportées ou saisies dans le même ordre. Une méthode de détection devrait être indépendante de l'ordre dans lequel les variables sont fournies.

2.2 ETUDES EMPIRIQUES : données utilisées

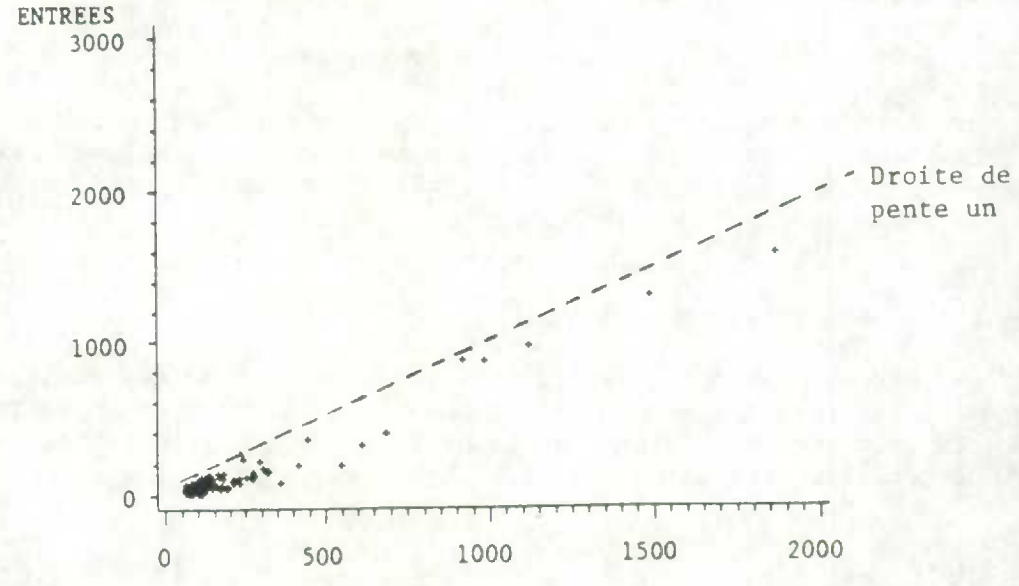
Les données finales des questionnaires longs du recensement des manufacturiers de 1985 ont été utilisées lors des études portant sur les deux méthodes de détection. Le nombre de dossiers étudiés se limite toutefois à ceux de la province de Québec soit 2887. Les deux variables utilisées sont les entrées et les sorties manufacturières.

Le graphique suivant représente la variable "entrées manufacturières" en fonction de la variable "sorties manufacturières". Afin de mieux visualiser l'allure du graphique, ce dernier a été divisé en deux parties. Le premier graphique représente tous les dossiers dont la variable sorties manufacturières est plus petite ou égale à soixante millions (95 % des dossiers). Le second graphique représente le reste des dossiers.

ENTREES vs SORTIES
(dossiers avec sorties <= $ 60 millions)
(Graphique en million de $)



ENTREES vs SORTIES
(dossiers avec sorties > $ 60 millions)
(Graphique en million de $)



En observant ces graphiques, on remarque que la variabilité augmente avec la valeur des variables. Le nuage de points est en forme d'entonnoir. On constate également qu'il existe une borne dans la partie supérieure du premier graphique, au-delà de laquelle il y a très peu de dossiers. Cette situation s'explique par une contrainte naturelle, à savoir que les entrées sont, sauf exception, toujours plus petites que les sorties. On observe de plus, que cette borne naturelle équivaut à une droite de pente un.

2.3 RÉGRESSION LINÉAIRE SIMPLE

2.3.1 Description de la méthode

Une des premières méthodes testées est celle de la régression linéaire simple. Les liens incohérents entre deux variables sont identifiés à l'aide d'un intervalle de confiance à 95 % calculé autour de la droite de régression. Tous les dossiers se trouvant en dehors de cette intervalle de confiance sont considérés comme suspects.

Il est bien connu que la régression est une méthode statistique qui est influencée par les valeurs extrêmes. Ainsi l'intervalle de confiance autour d'une droite de régression s'éloigne de la droite au fur et à mesure que les valeurs augmentent. On s'attend donc, suite aux propriétés énoncées à ce que la régression soit appropriée comme méthode de détection, puisque l'intervalle de confiance tendra à épouser la forme entonnoir des données.

Les données finales du RM 85 (Québec) ont été modèlisées et les résultats ont été appliqués sur le même ensemble de données. Lors de l'application des procédures de vérification, les données du cycle précédent seront modèlisées et les résultats seront appliqués sur les données du cycle courant.

Le modèle de régression ajusté est du type suivant :

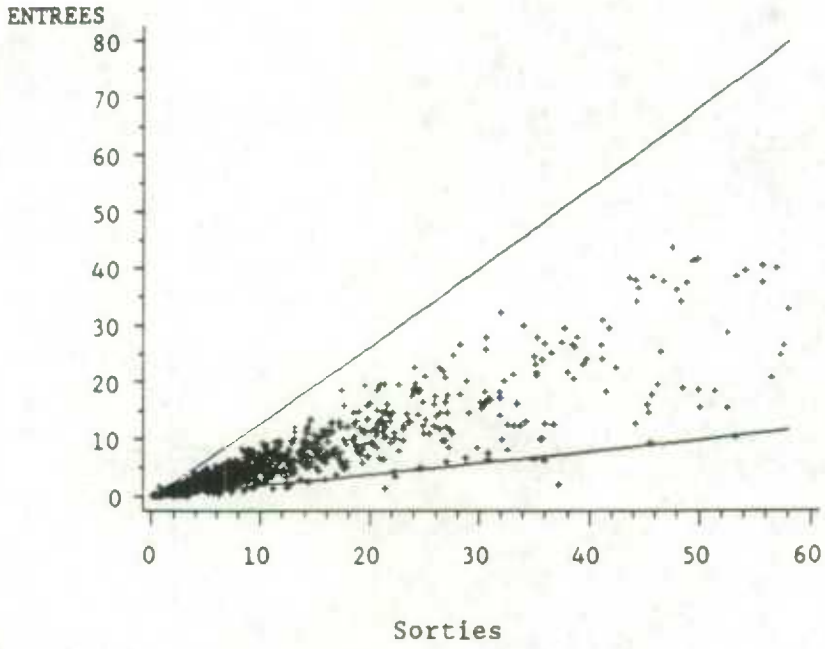
$$\text{Log (entrées)} = \alpha + \beta \log (\text{sorties}) + e$$

La fonction log a été utilisé afin d'éliminer la variabilité résiduelle. Les détails relatifs au modèle ainsi qu'à l'analyse résiduelle sont présentés à l'appendice 5. Les résultats obtenus confirment la validité du modèle ajusté.

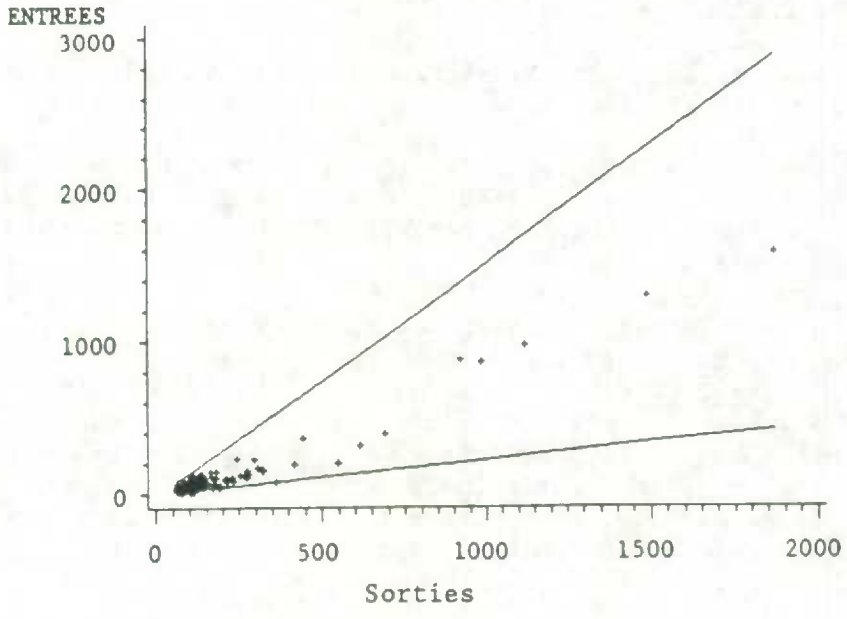
2.3.2 Résultats

La vitesse avec laquelle l'intervalle de confiance s'éloigne de la droite de régression a été sous-estimée. Les graphiques suivants représentent l'intervalle de confiance de la droite de régression projeté dans le plan des données originales (non-transformées).

INTERVALLE DE CONFIANCE DE LA DROITE DE RÉGRESSION
(dossiers avec sorties <= $ 60 millions)
(Graphique en million de $)



INTERVALLE DE CONFIANCE DE LA DROITE DE RÉGRESSION
(dossiers avec sorties > $ 60 millions)
(Graphique en million de $)



108 dossiers ont été rejetés par l'application de la méthode de la régression linéaire. On remarque aisément que l'intervalle de confiance n'épouse pas la forme des données. Au contraire, plus les valeurs sont grandes plus il s'en éloigne. On peut donc déduire que la majorité des dossiers rejetés par cette méthode sont des dossiers pour lesquelles les deux variables ont de petites valeurs.

2.3.3 Conclusions

La régression linéaire simple est trop influencée par les valeurs extrêmes. L'intervalle de confiance n'épouse pas la forme des données; plus les valeurs sont grandes plus l'intervalle s'élargit. La méthode de régression n'est, par conséquent, pas assez sévère pour les grandes valeurs.

De plus, les résultats diffèrent selon l'ordre dans lequel les variables sont traitées. En effet, le modèle de régression $x=\alpha+\beta y$ n'identifiera pas les mêmes dossiers suspects que le modèle $y=\alpha+\beta x$. Les résultats de la méthode de détection par régression linéaire dépendent donc de l'ordre des variables.

Par conséquent, la régression linaire simple n'est pas appropriée comme méthode de détection puisqu'elle ne rencontre pas les critères recherchés. Toutefois, d'autres avenues peuvent être explorées, par exemple, la régression non-paramétrique, la régression avec variable indicatrice ou encore l'utilisation d'une méthode différente de celle de l'intervalle de confiance pour identifier les dossiers suspects. Ces avenues seront explorées dans l'avenir si le temps et les budgets le permettent.

2.4 MÉTHODE DES TENDANCES

La méthode identifie les dossiers dont la tendance (relation) entre deux variables diffère significativement de la tendance globale correspondante des autres dossiers. Pour ce faire, elle tente de modéliser la tendance entre ces deux variables. Cette dernière peut se définir de deux façons : soit un rapport entre deux variables d'un même cycle d'enquête ou soit une différence relative entre une variable d'un cycle courant et la même variable du cycle précédent.

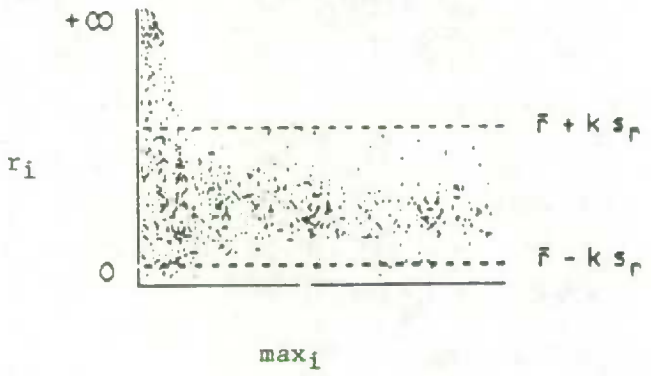
L'enquête mensuelle sur les manufacturiers (CSIO) utilise présentement, avec succès, la méthode des tendances pour le système de vérification dans les bureaux régionaux. La méthode est utilisée pour contrôler la différence relative entre une variable du mois courant et la variable correspondante du mois précédent. La méthode ayant fait ses preuves pour contrôler la relation entre deux cycles d'une enquête, le but de la présente étude est donc de vérifier l'efficacité de la méthode à contrôler le rapport entre deux variables du même cycle.

2.4.1 Description de la méthode

Le but de cette section est de fournir la théorie sous-jacente à la méthode des tendances [4]. Cette méthode tente de modéliser le rapport entre deux variables. L'allure du nuage de point du rapport entre deux variables (r_i) en fonction du maximum entre les deux variables est représentée par le graphique suivant.

$r_i = x_i / y_i \qquad max_i = max(x_i, y_i)$

ALLURE DU GRAPHIQUE DES r_i EN FONCTION DU MAXIMUM DES DEUX VARIABLES



Le graphique n'est pas symétrique. Il y a une compression vers zéro. De plus, la variabilité des rapports pour les petits dossiers est plus grande que la variabilité des rapports pour les grands.

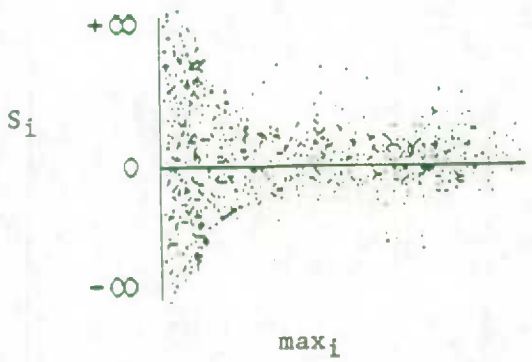
Si on appliquait des bornes du genre $r \pm ks_r$ (où r est la moyenne des rapports et s_r l'écart type des rapports) les résultats se traduiraient par une sur-représentation des petits dossiers et une sous-représentation des grands dossiers. C'est ce qu'on appelle l'effet de masque par la taille. Cet effet de masque est dû à la différence de variabilité entre les petits et les grands dossiers. Il se peut aussi, dans des circonstances particulières, que cette méthode ne permette pas de déceler des dossiers suspects dans la partie inférieure du graphique (borne inférieure négative) [7].

Pour remédier à cette situation, les tendances r_i doivent être transformées de façon à pouvoir déceler les dossiers suspects aux deux extrémités. Il faut rendre le graphique symétrique par rapport à zéro. Une des transformations possibles se définie comme suit :

$$S_i = \begin{cases} 1 - r_M/r_i & \text{si } 0 < r_i < r_M \\ r_i/r_M - 1 & \text{si } r_i \geq r_M \end{cases}$$

où r_M = médiane des r_i

ALLURE DU GRAPHIQUE DES S_i EN FONCTION DU MAXIMUM DES DEUX VARIABLES



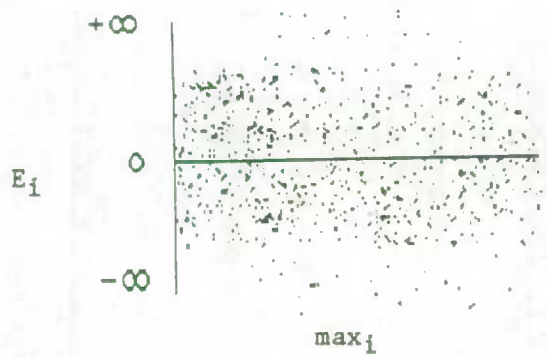
Cependant l'effet de masque par la taille est toujours présent. Si l'on veut tenir compte de la valeur des données, il est alors nécessaire de définir une transformation du type suivant :

$$E_i = S_i * [\text{Max } (x_i, y_i)]^u$$

où $0 < U <= 1$.

Les valeurs de E_i représentent des effets et l'exposant U détermine l'importance de la valeur des données. Cette transformation permet d'accorder plus d'importance à une petite variation touchant un grand dossier qu'à une grande variation touchant un petit dossier.

ALLURE DU GRAPHIQUE DES E_i EN FONCTION DU MAXIMUM DES DEUX VARIABLES



En désignant respectivement par E_{Q1}, E_M et E_{Q3} le premier quartile, la médiane et le troisième quartile des E_i, l'intervalle d'acceptation est définie de la façon suivante :

$$\text{Borne inférieure} = E_M - c*(E_M - E_{Q1})$$
$$\text{Borne supérieure} = E_M + c*(E_{Q3} - E_M)$$

Le paramètre c détermine la largeur de l'intervalle d'acceptation. Le paramètre U détermine la courbure des deux bornes. Tous les dossiers dont l'effet E_i correspondant se situe à l'extérieur de ces bornes sont définis comme étant suspects.

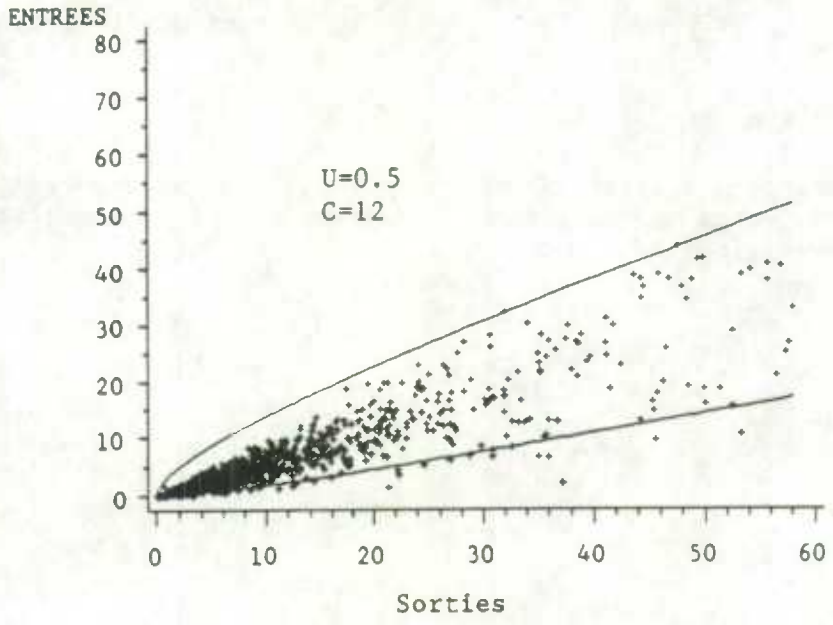
2.4.2 Résultats

Pour cette étude, on a utilisé le même ensemble de données (Québec 85) et les mêmes variables (entrées et sorties manufacturières) que lors de l'étude sur la régression linéaire simple.

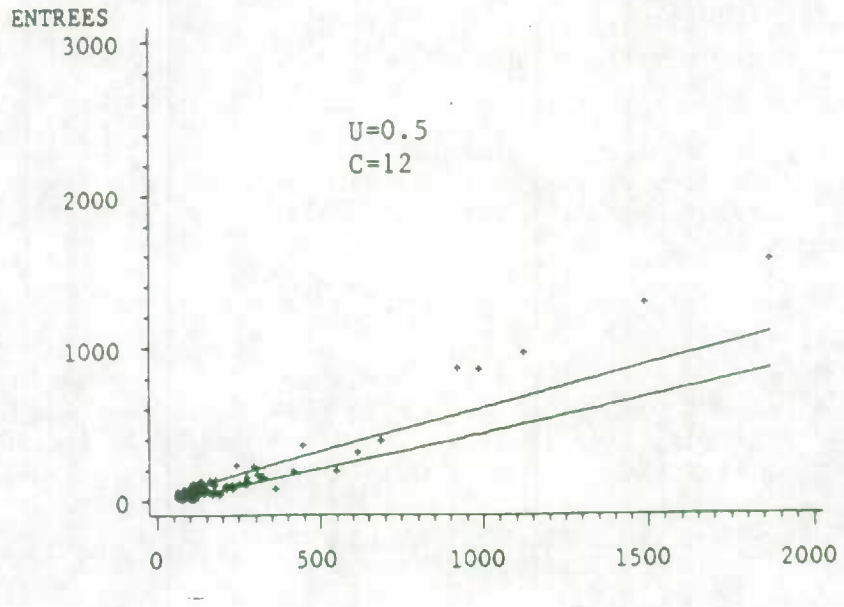
Le rapport étudié : $\frac{\text{entrées manufacturières}}{\text{sorties manufacturières}}$

Les données finales du RM 85 (Québec) ont été modèlisées et les résultats ont été appliqués sur le même ensemble de données. Lors de l'application des procédures de vérification, les données du cycle précédent seront modèlisées et les résultats seront appliqués sur les données du cycle courant. Le graphique suivant représente, dans le plan des données originales, l'intervalle d'acceptation obtenu par la méthode des tendances.

INTERVALLE D'ACCEPTATION DE LA MÉTHODE DES TENDANCES
(dossier avec sorties <= $ 60 millions)
(Graphique en million de $)



INTERVALLE D'ACCEPTATION DE LA MÉTHODE DES TENDANCES
(dossier avec sorties > $ 60 millions)
(Graphique en million de $)

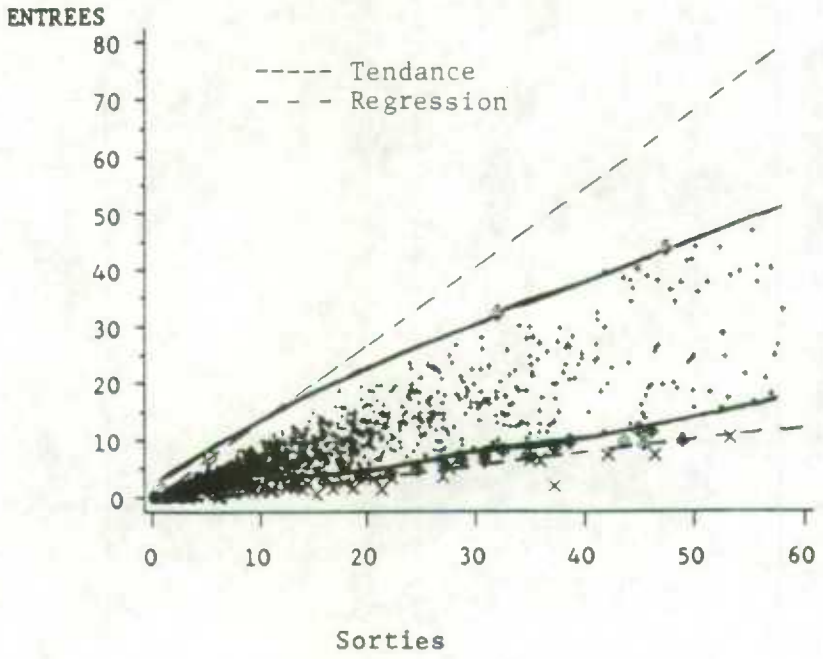


133 dossiers ont été rejetés par la méthode des tendances. On observe que les deux bornes épousent bien la forme des données et qu'elles semblent même se resserrer sur les grandes valeurs. La méthode a donc tendance à être plus sévère pour ces valeurs. On remarque également que très peu de dossiers se situant dans la partie supérieure du graphique ont été rejetés. Cette situation s'explique par la contrainte naturelle mentionnée précédemment, à savoir que les entrées manufacturières sont, sauf exception, toujours plus petites que les sorties manufacturières.

2.4.3 Comparaison

Comparons maintenant les résultats obtenus par les deux méthodes. Les graphiques suivants représentent simultanément les résultats des deux méthodes de vérification.

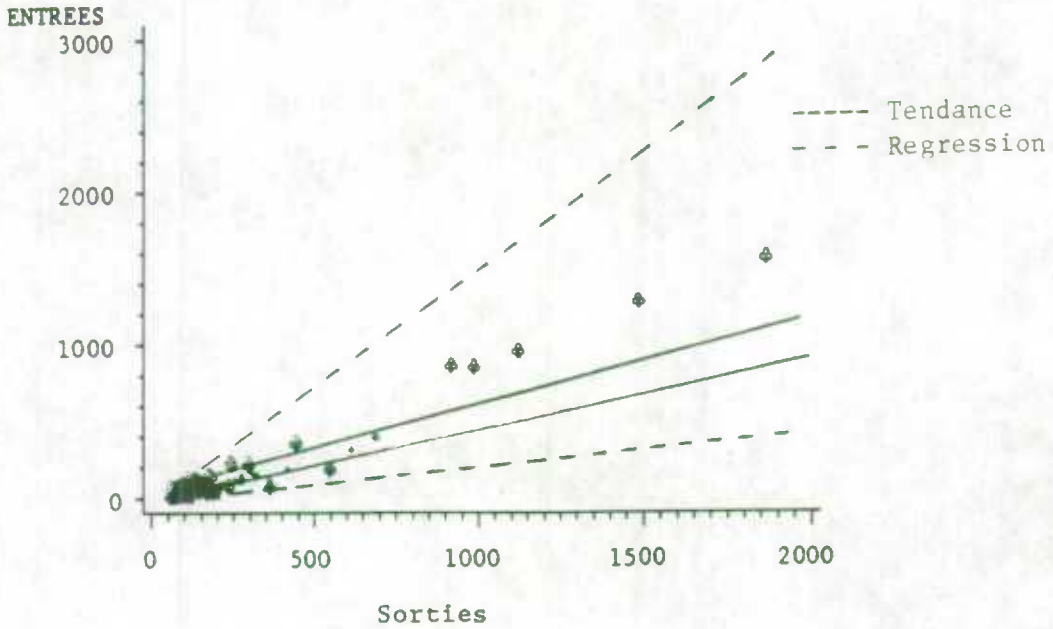
RÉSULTATS RÉGRESSION ET MÉTHODE DES TENDANCES
(dossiers avec sorties <= $ 60 millions)
(Graphique en million de $)



Légende

\+ : dossiers non-rejetés
O : dossiers rejetés seulement par la méthode de régression
◈ : dossiers rejetés seulement par la méthode des tendances
X : dossiers rejetés par les deux méthodes

RÉSULTATS RÉGRESSION ET MÉTHODE DES TENDANCES
(dossiers avec sorties > $ 60 millions)
(Graphqiue en million de $)



Légende

+ : dossiers non-rejetés
◆ : dossiers rejetés seulement par la méthode des tendances
X : dossiers rejetés par les deux méthodes

On constate immédiatement que l'écart entre les deux intervalles s'accentue. Plus les valeurs sont grandes plus les bornes de la régression s'éloigne des données alors qu'on observe le phénomène inverse pour la méthode des tendances.

Les dossiers rejetés par les deux méthodes ne sont pas tous les mêmes. Le tableau suivant indique le nombre de dossiers rejetés uniquement par la régression, le nombre de dossiers rejetés par les deux méthodes et finalement le nombre de dossiers rejetés uniquement par la méthode des tendances.

	Nb de dossiers Rejetés par	
Régression seulement	42	108 (régression)
Les deux méthodes	66	133 (tendances)
Tendances seulement	67	

En se référant aux graphiques précédents, il est facile de remarquer que la regression rejète des dossiers dont la valeur des deux variables à l'étude est plus petite. Des statistiques ont été compilées dans le but d'évaluer l'impact de chacune des méthodes sur le type de dossiers rejetés. Pour ce faire, on a mesuré, à quelle proportion du total des deux variables correspondent les dossiers rejetés par chacune des méthodes. Le tableau suivant présente les résultats obtenus.

Total des entrées manufacturières : 29 millions
Total des sorties manufacturières : 51 millions

	Nb de dossiers Rejetés	% du total Entrées	% du total Sorties
Régression seulement	42	.08 %	.20 %
	108	.76 %	2.99 %
Les deux méthodes	66	.68 %	2.79 %
	133	32.33 %	29.20 %
Tendances seulement	67	31.65 %	26.41 %
Tendances * seulement	59	10.42 %	11.15 %

* Calculs refaits en enlevant les huit dossiers ayant les plus grandes valeurs pour la variable sorties.

Les 42 dossiers rejetés uniquement par la régression correspondent à .08 % du total des entrées et à .2 % du total des sorties. Ces proportions sont très minimes si l'on compare avec les dossiers rejetés uniquement par la méthode des tendances qui eux correspondent à 31.65 % du total des entrées et à 26.41 % du total des sorties manufacturières.

Les résultats obtenus pour la méthode des tendances peuvent sembler "gonflés" puisqu'il est évident, lorsqu'on regarde les graphiques précédents que cette méthode identifie des dossiers dont les valeurs des deux variables sont extrêmement grandes. Les calculs ont donc été refaits, pour la méthode des tendances, en éliminant les huit dossiers ayant les plus grandes valeurs pour la variables sorties. Les résultats restent malgré tout concluants. Les 59 dossiers restants, correspondent respectivement à 10 % et 11 % du total des variable entrées et sorties manufacturières. Les dossiers identifiés comme suspects par la méthode des tendances sont beaucoup plus importants en terme d'impact sur les totaux que ceux identifiés par la régression.

2.4.4 Conclusions

Les analyses effectuées permettent de tirer les conclusion suivantes. Les bornes supérieure et inférieure épousent bien la forme du nuage de points et modèlisent adéquatement le comportemnt des dossiers avec des grandes valeurs. D'autre part, la méthode est indépendante de l'ordre dans lequel les variables sont fournies puisque les mêmes dossiers suspects seront identifiés que l'on travaille avec le rapport x/y ou y/x.

Cette méthode possède également des propriétés intéressantes. En effet, la distance entre les deux bornes ainsi que la forme de la courbe se contrôlent à l'aide de paramètres. La méthode offre ainsi la possibilité de maitriser, jusqu'à un certain point, le nombre de rejets. Chaque borne peut être ajustée séparément. Ceci donne le pouvoir de modéliser les dossiers se situant dans la partie supérieure du graphique différemment des autres. La méthode des tendances est aussi facile d'implantation. Il suffit de calculer à l'aide des données d'un cycle précédent, les bornes d'acceptation selon la classification désirée. Lors de la vérification, la tendance entre les deux variables à vérifier sera calculée puis comparée aux bornes déjà déterminées.

La méthode des tendances rencontre les critères recherchés. De plus elle identifie un nombre minimum de dossiers comme étant suspect. Ce qui implique que le nombre de confirmations/corrections à effectuer peut être minimum tout en conservant la qualité des données.

À titre d'exemple, les statistiques suivantes ont été prélevées lors d'une étude parallèle effectuée pour le compte de l'enquête mensuelle sur les manufacturiers (CSIO). Cette étude avait pour but de comparer l'ancienne méthode de détection (méthode manuelle) avec la méthode des tendances. Les résultats sont les suivants :

ENQUÊTE MENSUELLE SUR LES MANUFACTURIERS (CSIO)
Pourcentage de rejets
(Dossiers suspects nécessitant une confirmations/corrections)

	méthode manuelle	méthode tendances
Halifax	33.2 %	15.8 %
Winnipeg	41.9 %	14.1 %

On observe qu'en utilisant la méthode des tendances le nombre de rejets a été réduit de moitié. Ceci implique une réduction des coûts puisque le nombre de confirmations/corrections à effectuer a diminué. Cependant, on est en droit de se questionner sur la qualité des données. Hors, parmi les 131 dossiers effectivement corrigés par les interviewers utilisant l'ancienne méthode, 129 ont été identifiés par la méthode des tendances.

Avec la méthode des tendances, le nombre de dossiers suspects a diminué donc par conséquent les coûts reliés à la vérification sont aussi réduits. De plus, malgré cette réduction on conserve le même degré de qualité des données puisque la grande majorité des dossiers nécessitant vraiment une correction ont été identifiés. Il y a donc une grande possibilité d'économie au niveau des coûts et des ressources allouées au processus de vérification via l'utilisation d'une méthode telle que celle des tendances.

3. CONCLUSIONS

Cette section a pour but de récapituler les résultats importants obtenus au cours des recherches.

L'application de la méthode de classification a pour but d'identifier des groupes de variables afin de maximiser l'efficacité du processus de vérification. La méthode utilisée n'est pas la seule qui existe mais les résultats obtenus sont probants. Elle extrait facilement le patron de corrélation d'un ensemble de variable et forme des groupes cohérents dont les variables ont toutes un lien logique entre elles.

La méthode des tendances est une méthode performante. Elle rencontre tous les critères recherchés. Elle s'applique assi bien à la détection longitudinale (d'un cycle à l'autre d'une enquête) qu'à la détection à l'intérieur d'un même cycle d'enquête. De plus elle permet de réduire les coûts reliés au processus de vérification des données.

4. TRAVAIL FUTUR

Dans un avenir plus ou moins rapproché, divers travaux devront être effectués dans le but de compléter les recherches amorcées.

Développement

La tâche la plus importante à effectuer après avoir étudié et développé un ensemble de principes applicables à la vérification des données, est de fournir le support nécessaire à l'équipe de développement pour l'élaboration du système de production de la fonction générale de collecte et de saisie.

Travail non moins important, il faut aussi compléter le développement de la fonction de caractérisation. Cette fonction fait partie du processus intégré de vérification et de correction. Elle a pour but de classifier les dossiers suspects pour les besoins des procédures de suivi.

Ce document étant consacrer au traitement des variables quantitatives, on pourrait, dans le cadre de la fonction générale de collecte et de saisie, développer une approche pour les variables qualitatives.

Simulation

Plusieurs principes ont été mis de l'avant lors du développement des systèmes généraux. Afin de mettre en pratique ces principes on simulera le processus de saisie, vérification et imputation à l'aide de données d'enquête. Les résultats obtenus seront évalués en terme de coûts et de qualité des données.

Méthode de classification

Lors des études empiriques concernant la méthode de classification des variables, on a utilisé les données des recensement des manufacturiers. On devrait étudier le comportement de cette méthode lorsqu'appliquée sur les données d'une autre enquêt

Méthode(s) de détection multivariée

La méthode de classification sert à identifier des groupes de variables pour la vérification. Pour l'instant seules les méthodes bivariées ont été étudiées. Il serait intéressant de pouvoir utiliser une méthode de détection qui tient compte de toutes les variables d'un groupe. La détection multivariée étant un domaine peu connu on projète d'amorcer des recherches dans cette direction.

5. BIBLIOGRAPHIE

[1] Berthelot, J.-M. (1989), Approche générale pour la sous-fonction de vérification et de correction des données, Statistique Canada, Division des méthodes d'enquêtes entreprises, Document de travail.

[2] Harman, H.H. (1976), MODERN FACTOR ANALYSIS, 3rd ed, University of Chicago Press, Ill. (487 p.)

[3] Harris, W. et Keiser, F. (1964), Oblique factor analystic solutions by orthogonal transformations, PSYCHOMETRIKA, vol 29, no 24, p. 347-362.

[4] Hidiroglou, M.A. et Berthelot, J.-M., (1986), Contrôle statistique et imputation dans les enquêtes-entreprises périodiques, TECHNIQUES D'ENQUÊTES , Vol 12, no 1, pp. 79-89.

[5] Holzinger, K.J. et Harman, H.H. (1942), FACTOR ANALYSIS, University of Chicago Press, Ill. Chicago, Ill., (417 p.).

[6] SAS institue, (1985), SAS USER'S GUIDE : STATISTICS, Version 5, North Carolina.

[7] Wilkinson, R.G., (1982), An outlier identification technique designed for the business finance annual survey, Statistique Canada

APPENDICE 1

Description théorique de la méthode de classification.

Cet appendice contient une description algébrique des étapes de la méthode de classification concernant l'analyse en composantes principales et la rotation orthoblique.

Notation

I : Matrice identité

R : Matrice de corrélation entre les variables

Q : Matrice des vecteurs propres de la matrice R

Λ : Matrice diagonale contenant les valeurs propres de la matrice R

A : Matrice des coefficients (α_{ij}) de l'analyse en composantes principales ("factor loadings")

C_j : $j^{ième}$ composante principale

T : Matrice de rotation orthogonale

S : Matrice de corrélation entre les variables et les axes

Φ : Matrice de corrélation entre les axes

P : Matrice des nouveaux coefficients ("factors loadings") obtenus après une transformation de la matrice A

D : Matrice diagonale définie positive

1) Analyse en composantes principales

Une des premières étapes de la méthode de classification est d'effectuer séparémemt, l'analyse en composantes principales (ACP) de la matrice de corrélation de chacun des groupes.

Selon la théorie de l'ACP, on peut trouver A ("factor loadings") de façon à ce que $R = A\,A'$

On sait que $R = Q\,\Lambda\,Q'$ décomposition de la matrice de corrélation en vecteurs et valeurs propres

si on pose $A = Q\,\Lambda^{1/2}$

on a que $A\,A' = Q\,\Lambda^{1/2}\,\Lambda^{1/2}\,Q'$

$= Q\,\Lambda\,Q' = R$

La proportion de variance expliquée par une composante principale (C_j) se calcule de la façon suivante :

Variance expliquée par $C_j = \Sigma_i \alpha^2_{ij} = \lambda_j$

(où λ_j = $j^{ième}$ valeur propre)

La proportion de variance expliquée $P_j = \lambda_j + \Sigma_i \lambda_i$

Si $P_1 < k_1$ ou $P_2 > k_2$ alors le groupe doit être séparé.
(k_1 et $k_2$sont des paramètres déterminés à l'avance).

2) Rotation orthoblique

Par la suite, un groupe est séparé en effectuant une variante de la méthode de rotation orthoblique proposée par Harman [3]. Cette méthode se nomme ainsi puisqu'on obtient une solution oblique par une transformation composée d'une matrice de rotation orthogonale et d'une matrice diagonale définie positive.

Avant d'expliquer la théorie sous-jacente à cette méthode, un bref aperçu des méthodes othogonale et oblique est présenté.

<u>Dans le cas orthogonal,</u> la matrice A (coefficients de l'ACP) est tranformée à l'aide d'une matrice de rotation orthogonale T (où T'T=TT'=I). On a alors AT=B, où B devient la nouvelle solution orthogonale.

Algébriquement, la corrélation entre une variable (z_i) et un axe (F_j) s'exprime de la façon suivante :

$S = A \Phi \quad (s_{ij} = \rho(z_i,F_j))$

$$\rho(z_i,F_j) = a_{i1}\,\rho(F_j,F_1) + a_{i2}\,\rho(F_j,F_2) + a_{i3}\,\rho(F_j,F_3) + \ldots + a_{ij}\,\rho(F_j,F_j) + \ldots + a_{in}\,\rho(F_j,F_n)$$

Les axes étant perpendiculaires, on a que

$$\rho(F_i,F_j) = 1 \text{ si } i=j \text{ (} \Rightarrow \Phi = I \Rightarrow S = A \text{ } s_{ij}=a_{ij}\text{)}$$
$$= 0 \text{ sinon}$$

Dans le cas oblique, la matrice A est transformée à l'aide d'une matrice de rotation construite de façon à obtenir une solution dont les axes ne sont pas perpendiculaires. On appelle P la matrice des nouveaux coefficients obtenus, pour ainsi les distinguer des anciens coefficients A. Lors d'une rotation oblique les axes sont corrélés entre eux ($\Phi \neq I$). On a donc que :

$S = P\Phi$

Cas orthoblique

La première étape consiste à appliquer la méthode quartimax [3] sur les deux premiers vecteurs propres afin d'obtenir la matrice de rotation orthogonale T ($TT'=T'T=I$). La méthode quartimax maximise une fonction de 4^e puissance des vecteurs propres afin de déterminer l'angle de rotation θ.

La matrice T se définit alors comme suit :

$T = t_{11} = \cos(\theta)$
$t_{12} = \sin(\theta)$
$t_{21} = -\sin(\theta)$
$t_{22} = \cos(\theta)$
$t_{ij} = 1$ pour $i=j$ $\quad i,j = 3,\ldots,v$
$t_{ij} = 0$ pour $i \neq j$ $\quad i,j = 3,\ldots,v$ (où v = nombre de variables)

Par la suite il faut trouver D^{-1} une matrice diagonale servant à normaliser l'expression $T' \Lambda T$ de façon à obtenir la matrice de corrélation Φ avec des 1 sur la diagonale.

Calcul de D^{-1}

On veut que $\Phi_{ij} = 1$ si $i=j$

on a que $\Phi = D^{-1} T' \Lambda T D^{-1}$

Posons diagonale $(T' \Lambda T) = D^2$

$\sigma^2 = D^2$
$\sigma = D$
$1/\sigma = D^{-1}$

$[\text{diagonale } (T' \Lambda T)]^{-1/2} = D^{-1}$

$$\longrightarrow \quad D^{-1} T' \Lambda T D^{-1} = \frac{\sigma_{ij}^2}{\sigma_i \sigma_j} \; ; \quad \text{si } i=j \longrightarrow \frac{\sigma_i^2}{\sigma_i \sigma_i} = 1$$

Une fois que T et D on été calculés, la solution oblique complète se définit comme suit :

$P = Q\ T\ D$ (nouveaux coefficients)

$\Phi = D^{-1}\ T'\ \Lambda\ T\ D^{-1}$ (matrice de corrélation entre les axes)

$S = Q\ \Lambda\ T\ D^{-1}$ (matrice de corrélation entre les variables et les axes)

Les s_{ij} sont utilisés pour associer chacune des variables soit au premier axe ou soit au deuxième axe. Les résultats de cette étape donnent deux groupes de variables disjoints. L'ensemble de la procédure sera appliqué de nouveau sur les groupes jusqu'à ce qu'ils respectent tous les critères imposés sur P_1 et P_2.

La procédure "VARCLUS" de SAS utilise une technique similaire pour effectuer la rotation orthoblique.

APPENDICE 2

Cet appendice contient un exemplaire du questionnaire standard. Les variables utilisées lors des études ainsi que leur numéro d'identification sont indiqués sur le questionnaire.

Statistique Canada Statistics Canada

Division de l'industrie

Enquête annuelle des manufactures, 1987

☐ If you prefer receiving this questionnaire in English, please check box and return to Industry Division, Statistics Canada, Ottawa, K1A 0T6 or telephone (613) 951-9340

Formule RM 5—3100—1.2

Dans toute correspondance se rapportant à ce questionnaire, prière d'indiquer les chiffres ci-dessous

Adresse postale (rectifier s'il y a lieu)

Emplacement de cet établissement (rectifier s'il y a lieu)

INSTRUCTIONS RELATIVES A LA DECLARATION ET AU MANDAT. Les questionnaires doivent être retournés dûment remplis dans les 60 jours suivant leur réception. Le Guide de déclaration ci-inclus a pour but de vous aider à remplir ce questionnaire. Il est numéroté de façon à correspondre aux articles du questionnaire. Conserver un exemplaire pour vos dossiers. Déclaration exigée en vertu de la Loi sur la statistique.

ENTENTES DE PARTAGE DES DONNEES

Pour alléger le fardeau de réponse et assurer des statistiques plus uniformes, Statistique Canada a conclu des ententes avec divers ministères et organismes publiques en vue d'un échange de données.

En vertu de l'article 10 de la Loi sur la statistique, avec les bureaux statistiques des provinces de Terre-Neuve, de la Nouvelle-Ecosse, du Nouveau-Brunswick, du Québec, de l'Ontario, du Manitoba, de la Saskatchewan, de l'Alberta et de la Colombie-Britannique, en ce qui a trait aux établissements situés à l'intérieur des limites de leur province respective. Les lois sur la statistique de ces provinces prévoient, à peu de chose près, les mêmes dispositions que la Loi fédérale sur la statistique en ce qui a trait à la confidentialité et les mêmes sanctions contre la divulgation des renseignements.

En vertu de l'article 11 de la Loi sur la statistique du Canada avec le ministère des Finances et du Tourisme de l'Ile-du-Prince-Édouard, pour tous les établissements situés dans la province. Les ententes conclues en vertu de l'article 11 ne s'appliqueront pas à vos déclarations pour l'enquête des manufactures de 1987 si un agent ou une personne autorisée de votre société signifie au statisticien en chef, par écrit, qu'il s'y oppose et envoie sa lettre accompagnée du questionnaire dûment rempli à la Division de l'industrie de Statistique Canada. Veuillez préciser les organismes ou ministères susmentionnés auxquels il ne faut pas communiquer les données.

CONFIDENTIALITE

La loi interdit à Statistique Canada de publier des statistiques recueillies au cours de cette enquête qui permettraient d'identifier une entreprise sans que celle-ci en ait donnée l'autorisation par écrit au préalable. Les données déclarées sur ce questionnaire resteront confidentielles, elles serviront exclusivement à des fins statistiques et elles seront publiées seulement de façon agrégée. Les dispositions de la Loi sur la statistique qui traitent de la confidentialité ne sont modifiées d'aucune façon par la Loi sur l'accès à l'information ou toute autre loi.

1.9 **ANNEE DE DECLARATION** — La déclaration doit porter sur la dernière année financière terminée entre le 1er avril 1987 et le 31 mars 1988.

du | | | | 1 | 9 | 8 | au | | | | 1 | 9 | 8 |

Jour Mois Année Jour Mois Année

1.3.1 Cet établissement a-t-il été actif au cours de l'année précisée à 1.9 ci-dessus? Oui 1 ☐ Non 2 ☐

Si "Non", expliquer brièvement la situation _____

_____ remplir l'attestation ci-dessous et nous retourner ce questionnaire.

Date

1.3.2 Cet établissement a-t-il abandonné les affaires pendant l'année? Oui 1 ☐ _____ Non 2 ☐

1.3.3 Cet établissement a-t-il changé de propriétaire pendant l'année? Oui 1 ☐ _____ Non 2 ☐

Si "Oui", fournir les données pour l'année entière de déclaration. Si c'est impossible, produire les données pour la durée de vos activités et donner le nom et adresse de la personne à rejoindre pour obtenir le reste des renseignements.

Nom _____ Adresse _____

1.6.1 Forme juridique (cocher une case)

1 ☐ Particulier
2 ☐ Société en nom collectif
3 ☐ Entreprise constituée en corporation
4 ☐ Coopérative

Date

1.6.2 La forme juridique indiquée à 1.6.1 diffère-t-elle de ce qui figure dans votre dernière déclaration? Oui 1 ☐ _____ Non 2 ☐

ATTESTATION — J'atteste que les renseignements fournis ici sont, autant que je sache, complets et exacts.

Signataire autorisé	Fonction officielle	Date

Nom de la personne à rejoindre (en caractères d'imprimerie s.v.p.)	Téléphone		
	Indicatif régional	Numéro	Poste

Adresse, y compris le code postal (si elle diffère de celle ci-dessus)	Code postal	Télex

STANDARD TB/ CT — REG. B100 000

Canada

1.7 NATURE DE L'ENTREPRISE (décrire brièvement)

1.7.1 _____

1.7.2 Cela représente-t-il un changement avec l'année dernière? Oui 1 ☐ Non 2 ☐

1.8 SIEGES SOCIAUX ET UNITES AUXILIAIRES D'ENTREPRISES A ETABLISSEMENTS MULTIPLES

1.8.1 Cet établissement possède-t-il un siège social ou un bureau de direction * canadien dont les activités peuvent faire l'objet d'une déclaration distincte? Oui 1 ☐ Non 2 ☐ Si "Oui", indiquer le nom et l'adresse

Nom _____ Adresse _____

1.8.5 Cet établissement est-il servi par des unités auxiliaires * qui desservent également un ou d'autres établissements de votre entreprise? Oui 1 ☐ Non 2 ☐

* Les données de ce siège ou bureau ne doivent pas figurer ici

		Stocks de la période couverte par la présente déclaration	
		Stocks d'ouverture $ canadiens (omettre les cents)	Stocks de fermeture $ canadiens (omettre les cents)
2	**STOCKS** selon la valeur comptable, y compris ceux en transit ou en consignation au Canada (voir l'article 2 du Guide de déclaration)		
2.1.1	Ces chiffres comprennent-ils les stocks détenus mais n'appartenant pas à l'établissement? Oui 1 ☐ Non 2 ☐		
	Stocks manufacturiers		
2.1.2	Stocks de combustible	2.1.21	2.1.22
2.1.3	Stocks de matières premières, composantes et fournitures achetées	2.1.31	2.1.32
2.1.5	Stocks de produits en cours de fabrication	2.1.51	2.1.52
2.1.6	Stocks de produits finis	2.1.61	2.1.62
	Stocks non-manufacturiers		
2.2	Stocks de produits achetés pour la revente en l'état	2.2.01	2.2.02
2.3	Autres stocks non manufacturiers (préciser)	2.3.01	2.3.02
2.5	**Total des stocks de cet établissement**	2.5.01	2.5.02

3 **COMMANDES NON REMPLIES** (voir l'article 3 du Guide de déclaration) 3.1.0

3.1 Prière d'en déclarer la valeur (ou votre estimation la plus juste) au 31 décembre 1987

3.2 Avez-vous normalement un certain nombre de commandes non remplies (à l'exception des retards de livraison)? Oui 1 ☐ Non 2 ☐

5. **CONSOMMATION DE COMBUSTIBLE ET D'ELECTRICITE ACHETES** (voir l'article 5 du Guide de déclaration)

MODE D'EVALUATION DU COMBUSTIBLE ET DE L'ELECTRICITE ACHETES

5.0.2 Déclarez-vous la consommation telle que demandée? Oui 1 ☐ Non 2 ☐

		Indicatif des marchandises à l'usage de Statistique Canada	Coût à l'établissement $ canadiens (omettre les cents)
5.1	Charbon	261 6	
5.2	Gaz naturel	263 1	
5.3	Essence pour moteurs (sauf l'essence d'aviation)	431 2	
5.4	Kérosène, mazout pour poêles (mazout no 1)	432 2	
5.5	Huiles diesel	432 3	
5.6	Mazouts légers (nos 2 et 3)	432 4	
5.7	Mazouts lourds (nos 4, 5 et 6)	432 5	
5.8	Gaz de pétrole liquéfiés (propane, butane, etc.)	436 1	
5.9	Electricité achetée (y compris le coût de service)	497 1	
5.10	Vapeur	497 2	
5.10.1	Autres combustibles achetés et consommés (y compris l'essence d'aviation, etc.) (préciser)		
5.11	**Total du combustible et de l'électricité**		5.11.0

STANDARD

CERTAINES ENTRÉES DE L'ACTIVITÉ MANUFACTURIÈRE

6. **Matières premières, composantes, contenants, fournitures, etc. achetés et utilisés** dans les opérations manufacturières (voir l'article 6 du Guide de déclaration). – Ne pas inclure les matières, etc. produites par cet établissement pour son propre usage.

MODE D'ÉVALUATION DES MATIÈRES PREMIÈRES, ETC.	Indicatif des marchandises à l'usage de Statistique Canada	Unité de mesure	Quantité utilisée	Coût total à l'établissement $ canadiens (omettre les cents)
6.0.2 Déclarez-vous l'utilisation telle que demandée? Oui 1 ☐ Non 2 ☐				
6.1 Matières premières et composantes achetées et utilisées dans la fabrication (préciser les articles importants)				
6.2 **Total des articles de 6.1**				6.2.0

6.3 Contenants non restituables et autres matières et fournitures d'emballage et de livraison achetés et utilisés pour vos produits de propre fabrication (préciser les articles importants)	Réservé à Statistique Canada	
6.4 **Total des articles de 6.3**		6.4.0
6.6 Valeur totale des fournitures d'exploitation, d'entretien et de réparation achetées et utilisées dans la fabrication, sauf le combustible		6.6.0
6.7 **Total des matières premières, composantes, contenants, fournitures, etc. (6.2 + 6.4 + 6.6)**		6.7.0
6.8 Montant versé à d'autres établissements pour du travail exécuté sur des matières appartenant à cet établissement (voir l'article 6.8 du Guide de déclaration). Inclure les montants versés aux "employés à la pièce à l'extérieur", dans les industries du vêtement		6.8.0
6.9 **Total des matières premières, contenants, fournitures et montant versé pour du travail exécuté (somme de 6.7 + 6.8)**		6.9.0

CERTAINES ENTRÉES DE L'ACTIVITÉ NON MANUFACTURIÈRE	Coût total à l'établissement $ canadiens (omettre les cents)
7. Activités des établissements de commerce, de la construction, etc. (voir l'article 7 du Guide de déclaration)	
→ 7.1 Achats de produits d'autres établissements pour la revente dans l'état où ils ont été achetés (inclure les transferts de tels produits reçus d'autres établissements de votre société) (déclarer les ventes de ces produits à 9.1)	7.1.0
→ 7.2 Matières et fournitures achetées et utilisées dans vos travaux de construction nouvelle, exécutés par vos propres salariés, pour votre propre usage (seulement les articles imputables au compte des immobilisations et déclarés à 9.2)	7.2.0
→ 7.3 Matières et fournitures achetées et utilisées dans la production de machines et de matériel de tous genres, exécutée par vos propres salariés, pour votre propre usage (seulement les articles imputables au compte des immobilisation et déclarés à 9.3)	7.3.0
→ 7.4 Fournitures de bureaux achetées et utilisées	7.4.0
→ 7.5 Toutes autres matières et fournitures achetées et utilisées par votre établissement	7.5.0
→ 7.6 Total des articles de 7	7.6.0
→ 7.7 Total de certaines entrées de l'activité manufacturière et non manufacturière (6.9 + 7.6)	7.7.0

CERTAINES SORTIES DE L'ACTIVITÉ MANUFACTURIÈRE

8. Livraisons de produits de propre fabrication (voir l'article 8 du Guide de déclaration).

MODE D'ÉVALUATION DES LIVRAISONS			Indicatif des marchandises à l'usage de Statistique Canada	Unité de mesure	Quantité livrée	Valeur nette des livraisons, sauf la taxe de vente, les droits et taxes d'accise, les frais de livraison par des transporteurs à forfait ou publics, les escomptes et remises, etc. $ canadiens (omettre les cents)
8.0.2 Déclarez-vous les livraisons telles que demandées?	Oui 1 ☐	Non 2 ☐				
8.0.3 Les valeurs déclarées le sont-elles au prix final de vente?	Oui 1 ☐	Non 2 ☐				
Si la réponse est "non", veuillez préciser le mode d'évaluation						
8.0.4 _____		☐				
8.1 Produits livrés (préciser les produits importants)						
8.2 Valeur totale des livraisons de produits de propre fabrication (total de 8.1)						8.2.0

8. SORTIES DE L'ACTIVITÉ MANUFACTURIÈRE – fin

		Valeur nette des livraisons, sauf la taxe de vente, les droits et taxes d'accise, les frais de livraison par des transporteurs à forfait ou publics, les escomptes et remises, etc. $ canadiens (omettre les cents)
8.3	Moins les rectifications pour tenir compte des articles ci-dessous si vous n'avez pas pu les exclure de la valeur de chacun des produits mentionnés à 8.1	
8.3.1	Paiements globaux pour les frais de livraison par des transporteurs à forfait ou publics	8.3.1
8.3.2	Paiements globaux de taxes de vente et de droits et taxes d'accise	8.3.2
8.3.3	Total des escomptes, reprises et remises	8.3.3
8.4	Total des ajustements (somme des articles de 8.3)	8.4.0
	Si les montants déclarés ci-dessus comprennent certaines sommes dont il est fait mention à la question 9.1, produits achetés pour être revendus comme tels, prière de le signaler ci-après ☐	
8.5	Valeur rectifiée des livraisons de produits de propre fabrication (8.2 moins 8.4 ou 8.2 si 8.4 est zéro)	8.5.0
8.6	Montant reçu en paiement du travail exécuté sur des matières et des produits appartenant à d'autres établissements (inclure les montants reçus d'autres établissements de votre entreprise)	8.6.0
8.7	**Valeur totale des livraisons de produits de propre fabrication et montant reçu en paiement du travail exécuté (8.5 + 8.6)**	8.7.0

9. **CERTAINES SORTIES PROVENANT DE L'ACTIVITE NON MANUFACTURIERE** (voir l'article 9 du Guide de déclaration)

	$ canadiens (omettre les cents)
9.1 **Valeur totale des livraisons de produits vendus dans l'état où ils ont été achetés** (les achats de ces produits doivent figurer à 7.1)	9.1.0
Indiquer ci-dessous les principaux produits compris dans cette valeur et donner le pourcentage estimatif de cette valeur (9.1) que chacun représente Produits — Pourcentage estimatif	
9.2 **Valeur comptable de la construction nouvelle, exécutée par vos propres salariés, pour votre propre usage** (seulement le montant imputable au compte des immobilisations — doit tenir compte du coût des matériaux déclarés à 7.2 et de celui de la main-d'oeuvre à 14.1.4)	9.2.0
9.3 **Valeur comptable des machines et du matériel produits par vos propres salariés et pour votre propre usage** (seulement le montant imputable au compte des immobilisations — doit tenir compte du coût des matériaux déclarés à 7.3 et de celui de la main-d'oeuvre à 14.1.4)	9.3.0
9.5 **Recettes provenant du loyer ou de la location de machines et de matériel fabriqués par votre établissement**	9.5.0
9.6 **Toutes autres recettes provenant des produits et services** (à l'exception des recettes hors exploitation comme les intérêts, les dividendes et autres recettes de location)	9.6.0
9.7 **Total des postes de la rubrique 9**	9.7.0
10. **Total général de certaines sorties de l'activité manufacturière et non manufacturière** (8.7 + 9.7)	10.0.0
RENSEIGNEMENTS ADDITIONNELS	
11. **Recettes provenant du loyer ou de la location de biens immobiliers** (terres, bâtiments, bureaux, etc.)	11.0.0
12. **Recettes provenant du loyer ou de la location de machines et de matériel autres que celles qui sont déclarées à 9.5 ci-dessus** (c-à-d. de machines de tous genres, de moteurs, de camions de tous genres, de remorques, de tracteurs, d'autre outillage, etc.)	12.0.0

13. **EXPORTATIONS**

Veuillez estimer quel pourcentage des sorties (ligne 10, ci-haut) sont livrées

13.61 au Canada _____ %

13.62 aux Etats-Unis _____ %

13.63 à d'autres pays _____ %

Total 100%

EMPLOYEES OF THIS ESTABLISHMENT (refer to section 14 in the Reporting Guide)

	Gross salaries, wages, commissions, bonuses, etc. (omit cents)	Average number employed during reporting period		Number of person-hours (please provide reasonable estimate where records are not maintained)	
		Male	Female	Worked	Paid
14.1 **Employees at this location**					
14.1.1 Executive, administrative and sales staff	14.1.11	14.1.12	14.1.13	XXXXX	XXXXX
14.1.2 Employees in manufacturing operations	14.1.21	14.1.22	14.1.23	14.1.24	14.1.25
14.1.4 Other workers, including employees engaged in construction and production of machinery and equipment for own use (see 9.2 and 9.3)	14.1.41	14.1.42	14.1.43	XXXXX	XXXXX
14.1.9 Total employees at this location (14.1.1 + 14.1.2 + 14.1.4)	14.1.91	14.1.92	14.1.93	XXXXX	XXXXX
14.2 **Employees at other locations**					
14.2.1 Employees in manufacturing operations	14.2.11	14.2.12	14.2.13	14.2.14	14.2.15
14.2.2 All other employees	14.2.21	14.2.22	14.2.23	XXXXX	XXXXX
14.3 Total employees at other locations (give details in section 15)	14.3.01	14.3.02	14.3.03	14.3.04 XXXXX	14.3.05 XXXXX

	Employees in manufacturing operations (see 14.1.2)	Executive, administrative and sales staff (see 14.1.1)
14.4 Supplementary information		
14.4.1 Average hourly rate of pay in dollars and cents		XXXXX
14.4.2 Number of hours in standard work week		
14.4.3 Average paid vacation (number of weeks per year)		
OR		
14.4.4 Vacation pay as % of earnings		
14.4.5 Number of paid statutory holidays per year		
14.4.6 Did you have a program of work sharing or reduced hours during the year?	Yes 1☐ No 2☐	Yes 1☐ No 2☐

15. **Employees at other locations included in this return** (attach separate sheet if necessary)

Other locations (street and number, municipality name, province)	Major activity carried on	Statistics Canada use	Gross salaries, wages, commissions, bonuses, etc. (omit cents)	Average number employed during reporting period	
				Male	Female
Total (should agree with 14.3 above)					

For unincorporated firms only	Male	Female
	Number	
16. Working owners and partners (do not include in 14 above)		

STANDARD

Canada

APPENDICE 3

Étude sur les coefficients de corrélation.

La méthode de classification des variables utilise une matrice de corrélation en entrée. Des études ont été effectuées dans le but d'évaluer le comportement de tois coefficient de corrélation : Pearson, Spearman et Kendall.

Le coefficient de corrélation de Pearson =

$$\frac{\Sigma_i\ (x_i-x)\ (y_i-y)}{\sqrt{(\Sigma_i\ (x_i-x)^2\ \Sigma_i\ (y_i-y)^2)}}$$

Le coefficient de corrélation de Spearman =
(coefficient non-paramétrique basé sur les rangs) :

$$\frac{\Sigma_i\ (r_i-r)\ (s_i-s)}{\sqrt{(\Sigma_i\ (r_i-r)^2\ \Sigma_i\ (s_i-s)^2)}}$$

où r_i est le rang de la $i^{ème}$ valeur de x

et s_i est le rang de la $i^{ème}$ valeur de y

Le coefficient de corrélation de Kendall =
(coefficient non-paramétrique basé sur les signes) :

$$\frac{S}{n(n-1)+2}$$ où n = nombre d'observations

1) Ordonner les observations selon l'ordre croissant de la variable x

2) Comparer chaque y avec tous les y qui se trouvent en-dessous. Une paire de y (un y comparé à un y situé en dessous) sera dans l'ordre naturel si le y situé plus bas est plus grand que l'autre. Autrement une paire sera dans l'ordre naturel inverse.

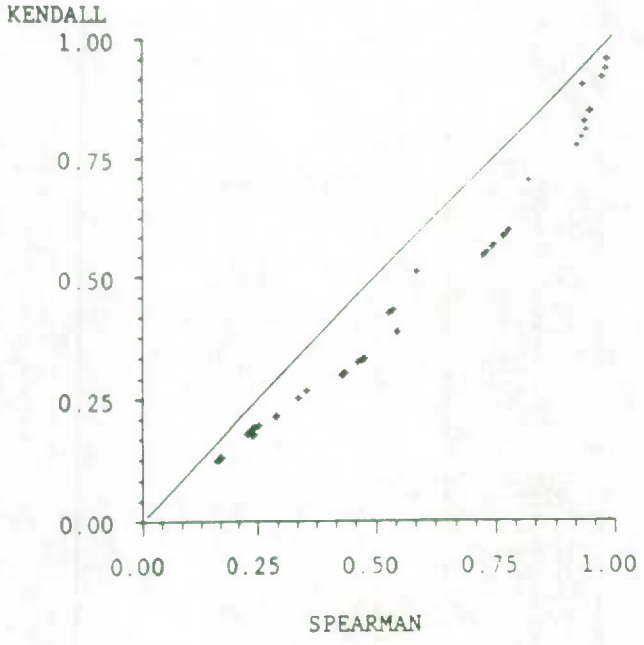
3) P est le nombre de paires dans l'ordre naturel Q est le nombre de paires dans l'ordre naturel inverse.

4) S=P-Q

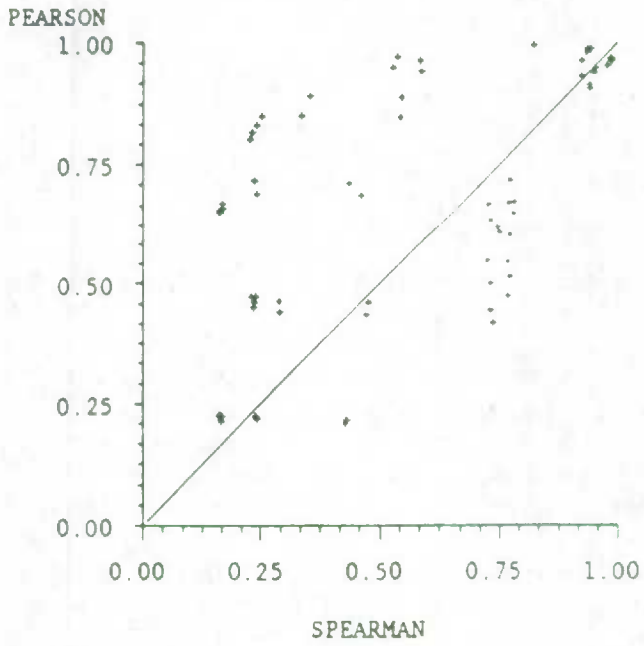
n(n-1)+2 = nombre total de paires

La matrice de corrélation a été calculée sur le même ensemble de données pour chacun des trois coefficients. Afin de comparer leur comportement on a tracé les graphiques représentant les valeurs d'un coefficient de corrélation en fonction de l'autre. Une droite de pente un a été ajoutée aux graphiques afin d'en faciliter l'interprétation.

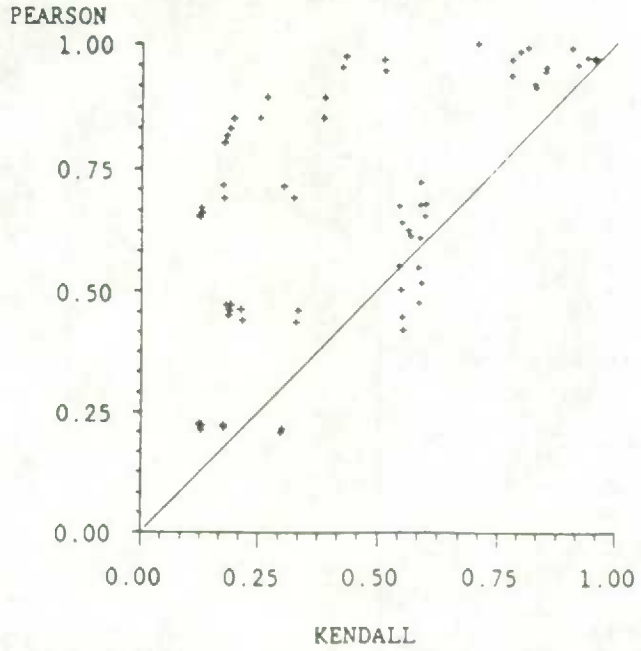
kENDALL vs SPEARMAN



PEARSON vs SPEARMAN



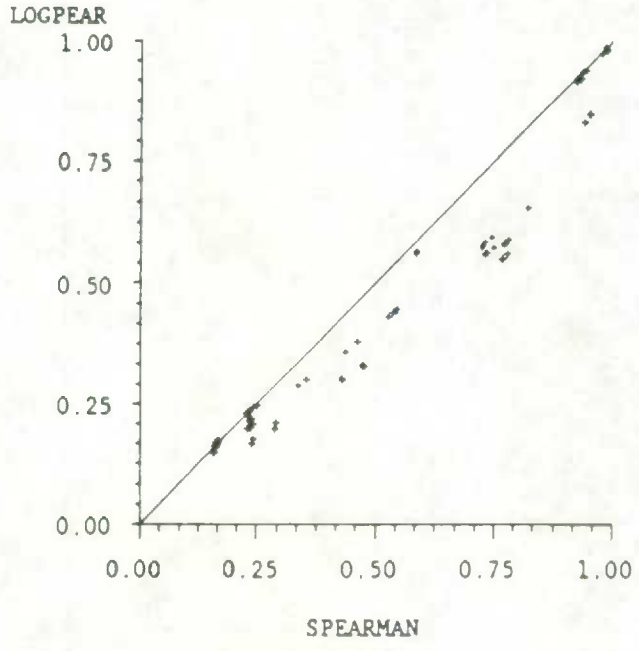
PEARSON vs KENDALL



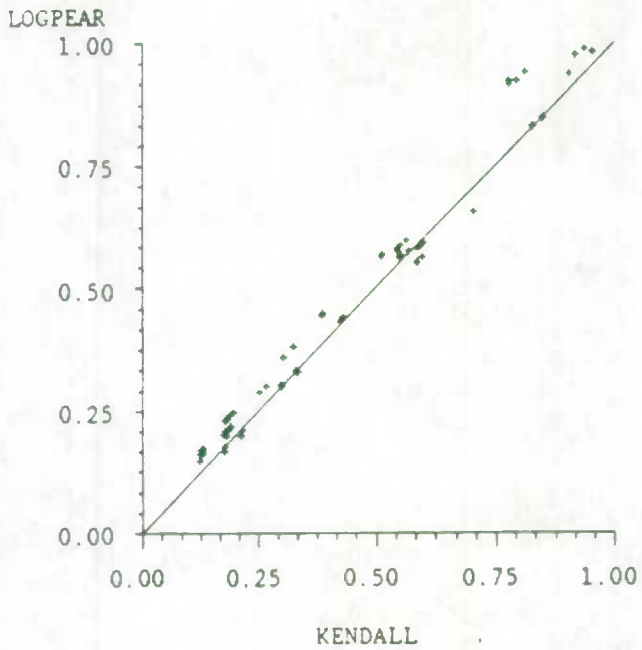
En analysant le premier graphique (Kendall vs Spearman) on constate que les deux coefficients ont un comportement similaire. Kendall semble toutefois être légèrement plus sévère (<) que Spearman. Les deuxième et troisième graphiques (Pearson vs Spearman et Pearson vs Kendall) se ressemblent. Pearson tend à être en général plus grand que les deux autres coefficients.

On a vu dans la section 1.3.2 qu'une transformation des données est nécessaire pour rendre Pearson mois dépendant des grandes valeurs. Les graphique deux et trois ont été refaits mais cette fois en utilisant la corrélation de Pearson calculée sur le logarithme des variables. Voici les résultats :

PEARSON (LOG) vs SPEARMAN



PEARSON (LOG) vs KENDALL



En transformant les données à l'aide du logarithme, on a diminué l'impact des grandes valeurs, le coefficient de corrélation de Pearson devient ainsi plus sévère. Son comportment est similaire à celui des deux coefficients non-paramétriques.

APPENDICE 4

Liste des groupes formés lors de la première étude empirique pour la méthode de classification

Résultats obtenus pour le recensement de 1985 (30 groupes). La lettre C représente le RM de 1985. (numéros des variables voir Annexe 2).

Groupe 1

C6.9.0
C7.7.0
C8.2.0
C8.5.0
C8.7.0
C10.0.0

Groupe 2

C14.2.11
C14.2.12
C14.2.13
C14.2.14
C14.2.15
C14.3.04
C14.3.05

Groupe 3

C5.11.0
C6.2.0
C6.7.0
C14.1.21
C14.1.22
C14.1.24
C14.1.25
C14.1.91
C14.1.92

Groupe 4

C14.2.21
C14.2.22
C14.2.23
C14.3.01
C14.3.02
C14.3.03

Groupe 5

C2.1.31
C2.1.32
C2.5.01
C2.5.02

Groupe 6

C7.1.0
C7.6.0
C9.1.0
C9.7.0

Groupe 7

C8.3.1
C8.3.3
C8.4.0

Groupe 8

C14.1.11
C14.1.12
C14.1.13

Groupe 9

C2.1.21
C2.1.22

Groupe 10

C2.1.51
C2.1.52

Groupe 11

C2.1.61
C2.1.62

Groupe 12

C2.2.01
C2.2.02

Groupe 13

C2.3.01
C2.3.02

Groupe 14

C7.2.0
C9.2.0

Groupe 15

C7.3.0
C9.3.0

Groupe 16

C14.1.23
C14.1.93

Groupe 17

C14.1.41
C14.1.42

Groupe 18

C3.1.0

Groupe 19

C6.4.0

Groupe 20

C6.6.0

Groupe 21

C6.8.0

Groupe 22

C7.4.0

Groupe 23

C7.5.0

Groupe 24

C8.3.2

Groupe 25

C8.6.0

Groupe 26

C9.5.0

Groupe 27

C9.6.0

Groupe 28

C11.0.0

Groupe 29

C12.0.0

Groupe 30

C14.1.43

Résultats obtenus pour le recensement de 1984 (28 groupes). La lettre Q représente le RM de 1984.

24 des 28 groupes formés pour le RM de 1984 sont exactement les mêmes que pour le RM de 1985. Les quatre groupes ayant changer sont présentés.

Groupe 1

Combinaison des groupes 1 et 20 de RM de 85

Q6.6.0
Q6.9.0
Q7.7.0
Q8.7.0
Q10.0.0

Groupe 2

Combinaison des groupes 2 et 4 du RM de 1985

Q14.2.11
Q14.2.12
Q14.2.13
Q14.2.14
Q14.2.15
Q14.2.21
Q14.2.22 Note : Q14.3.04 et Q14.3.05 n'exitaient pas lors du RM de 1984

Groupe 3

Combinaison des groupes 6 et 12 du RM de 1985

Q2.2.01
Q2.2.02
Q7.1.0
Q7.6.0
Q9.1.0
Q9.7.0

Groupe 4

Deux variables du groupe 1 du RM de 1985

Q8.2.0
Q8.5.0

Résultats obtenus lors de la classification des 146 variables des deux recensements. La lettre C représente le RM de 1985 et la lettre Q le RM de 1984.

Groupe 1

C6.9.0
C7.7.0
C8.7.0
C10.0.0
Q6.9.0
Q7.7.0
Q8.7.0
Q10.0.0

Groupe 2

C8.2.0
C8.5.0
Q8.2.0
Q8.5.0

Groupe 3

C14.2.11
C14.2.12
C14.2.14
C14.2.15
C14.3.04
C14.3.05
Q14.2.11
Q14.2.12
Q14.2.14
Q14.2.15
Q14.3.01
Q14.3.02

Groupe 4

C5.11.0
C14.1.21
C14.1.22
C14.1.24
C14.1.25
C14.1.91
C14.1.92
Q5.11.0
Q14.1.21
Q14.1.22
Q14.1.24
Q14.1.25
Q14.1.91
Q14.1.92

Groupe 5

C6.2.0
C6.7.0
Q6.2.0
Q6.7.0

Groupe 6

C14.2.21
C14.2.22
C14.2.23
C14.3.01
C14.3.02
C14.3.03

Groupe 7

Q14.2.21
Q14.2.22
Q14.2.23

Groupe 8

C2.1.31
C2.1.32
Q2.1.31
Q2.1.32

Groupe 9

C2.5.01
C2.5.02
Q2.5.01
Q2.5.02

Groupe 10

C7.1.0
C7.6.0
C9.1.0
C9.7.0
Q7.1.0
Q7.6.0
Q9.1.0
Q9.7.0

Groupe 11

C8.3.1
C8.4.0
Q8.3.1
Q8.4.0

Groupe 12

C8.3.3
Q8.3.3

Groupe 13

C14.1.11
C14.1.12
C14.1.13
Q14.1.11
Q14.1.12
Q14.1.13

Groupe 14

C2.1.21
C2.1.22
Q2.1.21
Q2.1.22

Groupe 15

C2.1.51
C2.1.52
Q2.1.51
Q2.1.52

Groupe 16

C2.1.61
C2.1.62
Q2.1.61
Q2.1.62

Groupe 17

C2.2.01
C2.2.02
Q2.2.01
Q2.2.02

Groupe 18

C2.3.01
C2.3.02
Q2.3.01
Q2.3.02

Groupe 19

C7.2.0
C9.2.0
Q7.2.0
Q9.2.0

Groupe 20

C7.3.0
C9.3.0
Q7.3.0
Q9.3.0

Groupe 21

C14.1.23
C14.1.93
Q14.1.23
Q14.1.93

Groupe 22

C14.1.41
C14.1.42
Q14.1.41
Q14.1.42

Groupe 23

C3.1.0
Q3.1.0

Groupe 24

C6.4.0
Q6.4.0

Groupe 25

C6.6.0
Q6.6.0

Groupe 26

C6.8.0
Q6.8.0

Groupe 27

C7.4.0
Q7.4.0

Groupe 28

C7.5.0
Q7.5.0

Groupe 29

C8.3.2
Q8.3.2

Groupe 30

C8.6.0
Q8.6.0

Groupe 31

C9.5.0
Q9.5.0

Groupe 32

C9.6.0
Q9.6.0

Groupe 33

C12.0.0
Q12.0.0

Groupe 34

C14.1.43
Q14.1.43

Groupe 35

C11.0.0
Q14.1.93

Groupe 36

C14.2.13
Q14.2.13
Q14.3.03

APPENDICE 5

Analyse du modèle de régression linéaire ajusté sur les données du recensement des manufacturiers de 1985 (Québec).

Modèle :

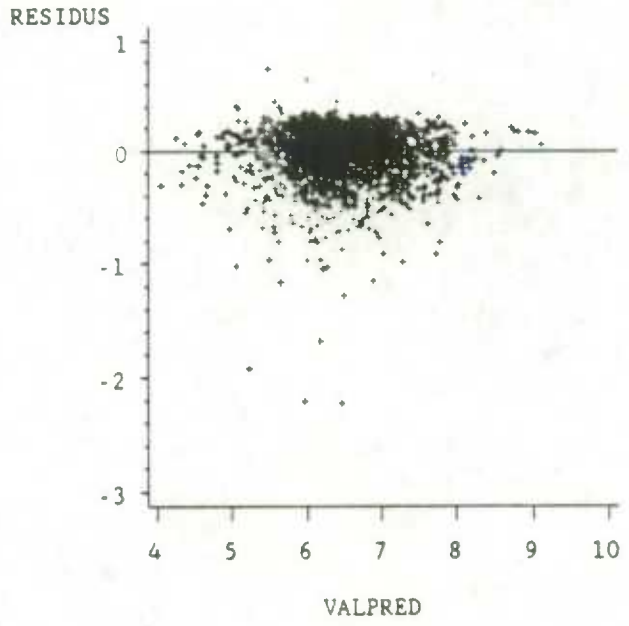
Log(entrées) = -.54 + 1.03 log (sorties) + e

R carré = .88

F = 21213.79

Il existe plusieurs façon de vérifier l'homocédasticité (égalité des variances) des résidus dans un modèle de régression. Une des plus simple est de tracer le graphique des résidus en fonction des valeurs prédites et d'en examiner la forme.

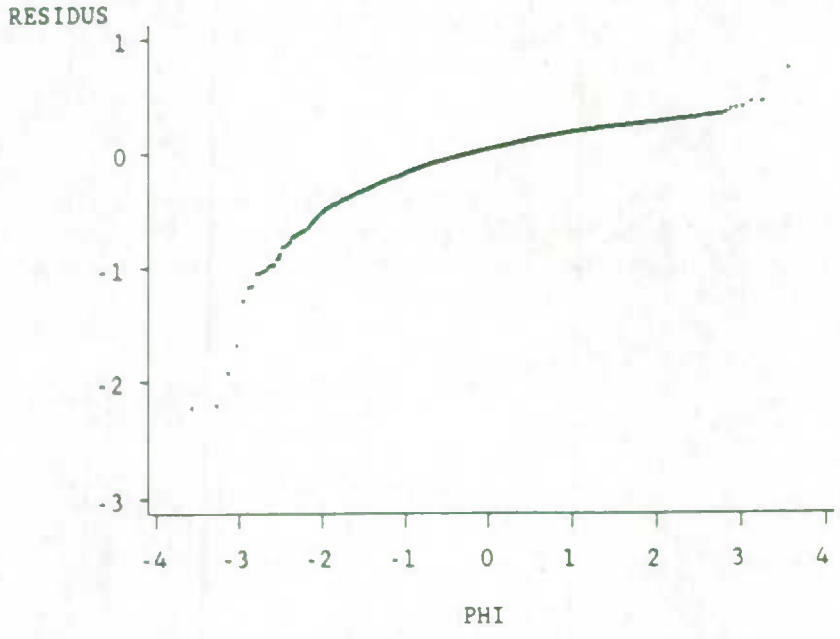
GRAPHIQUE DES RÉSIDUS EN FONCTION DES VALEURS PRÉDITES



Face au résultat obtenu, on peut supposer l'égalité des variances des résidus. Le graphique prend la forme d'un nuage de points distribués de façon aléatoire autour d'une droite de pente nulle et il ne présente aucun patron systématique. Les résidus semblent toutefois être bornés supérieurement. Ceci s'explique encore une fois par la contrainte naturelle qui existe entre les entrées et les sorties manufacturières.

Pour qu'un modèle soit adéquat, il faut que les résidus soit distribués selon une loi normale. La droite de Henri a été tracée pour vérifier cette hypothèse.

DROITE DE HENRI À PARTIR DES RÉSIDUS



Le graphique prend la forme d'une droite sauf aux deux extrémités. On peut donc supposer la normalité des résidus. Le modèle ajusté aux variables entrées et sorties manufacturières est donc adéquat et servira à la détection des dossiers suspects.