AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

FACULTY OF APPLIED ECONOMICS

DEPARTMENT OF MATHEMATICS,
STATISTICS AND ACTUARIAL SCIENCES

French Regional Wheat Prices : 1756-1870
An Illustrated Statistical Analysis

E. Borghers

RESEARCH PAPER 2006-003
February 2006

University of Antwerp, Prinsstraat 13, B-2000 ANTWERP, Belgium
Research Administration - Room B.213
phone: (32) 3 220 40 32
e-mail: joeri.nys@ua.ac.be

The papers can be also found at our website:
www.ua.ac.be/tew
(research > working papers)

D/2006/1169/003

French Regional Wheat Prices : 1756-1870
An Illustrated Statistical Analysis

S'ils n'ont pas de pain, qu'ils mangent des brioches
Marie-Antoinette

Prof. dr. E. Borghers

University of Antwerp
Faculty of Applied Economics
Department of Mathematics, Statistics and Actuarial Sciences

Key Words : Multivariate Statistics, Wheat Prices, France

Acknowledgement

The helpful comments and suggestions made by dr. K. Van Rompay are greatly appreciated.

Corresponding Address

University of Antwerp
Faculty of Applied Economics
Room C 442
Prinsstraat 13
B 2000 Antwerp - Belgium

eddy.borghers@ua.ac.be
T +32 (0) 3 220 41 31 (Office)
T +32 (0) 3 220 41 54 (Secretary)
F +32 (0) 3 220 48 57 (Secretary)

Contents

Section 1	Introduction	5
Section 2	Box-and-Whisker Plot	8
Section 3	Correlation Analysis	11
Section 4	Principal Component Analysis - Component Loadings	16
Section 5	Cluster Analysis - Original Data	20
Section 6	Cluster Analysis - Original Data Transposed	26
Section 7	Multidimensional Scaling Analysis	29
Section 8	Final Conclusions and Remarks	32
Appendix 1	Data	33
Appendix 2	Grands Secteurs Territoriaux	34
References		35

Table 1	Nine Regions - Grands Secteurs Teritoriaux	6
Table 2	Wheat Price for Nine Regions - Period : 1756-1870 Correlation Matrices	12
Table 3	Wheat Price for Nine Regions - Period : 1756-1870 Two-Component Model - Component Loadings - Varimax Rotation	18
Table 4	Wheat Price for Nine Regions - Period : 1756-1870 Tree Clustering - Ward's Method - Composition of Clusters	21
Table 5	Wheat Price for Nine Regions - Period : 1756-1870 Cluster Members and Distances from Cluster Means	22
Table 6	Wheat Price for Nine Regions - Period : 1756-1870 Euclidean Distances and Pearson Correlation between Cluster 1 and Cluster 2	23
Table 7	Wheat Price for Nine Regions - Period : 1756-1870 Transposed Data Sets - 3-Means Clustering - Cluster Members	26
Table 8	Wheat Price for Nine Regions - Period : 1756-1870 Transposed Data Sets - 3-Means Clustering - Cluster Members & Cluster Means	27
Table 9	Wheat Price for Nine Regions - Period : 1756-1870 MDS - Summary of Numerical Results - Goodness-of-Fit Tests	30

Figure 1	General Aggregated Wheat Price for France - Period : 1726-1913 Sequential Line Plot - Distance Weighted Least Squares Fit	5
Figure 2	Wheat Price for Nine Regions - Period : 1756-1870 Sequential Line Plot - Distance Weighted Least Squares Fit	6
Figure 3	Wheat Price for Nine Regions - Period : 1756-1870 Multiple Box-and-Whisker Plots - Connected Medians	9
Figure 4	Wheat Price for Nine Regions - Period : 1756-1870 Matrix Plots - Histogram - Scatterplot - Distance Weighted Least Squares Fit	13
Figure 5	Wheat Price for Nine Regions - Period : 1756-1870 Correlations between Center and Outer Regions	14
Figure 6	Wheat Price for Nine Regions - Period : 1756-1870 Correlations among Outer Regions	15
Figure 7	Wheat Price for Nine Regions - Period : 1756-1870 Largest Eigenvalues - Scree Plot	17
Figure 8	Wheat Price for Nine Regions - Period : 1756-1870 Two-Component Model - Varimax Rotated Loadings - Scatter Plot	19
Figure 9	Wheat Price for Nine Regions - Period : 1756-1870 Tree Clustering - Ward's Method - Structure of Clusters	21
Figure 10	Wheat Price for Nine Regions - Period : 1756-1870 Cluster Means - Sequential Line Plot - Distance Weighted Least Squares Fit	24

Figure 11	Wheat Price for Nine Regions - Period : 1756-1870 Original Data Sets - Cluster Means and Aggregated Price Variable Sequential Line Plot - Distance Weighted Least Squares Fit	25
Figure 12	Wheat Price for Nine Regions - Period : 1756-1870 Transposed Data Sets - 3-Means Clustering - Cluster Means and General Mean	28
Figure 13	Wheat Price for Nine Regions - Period : 1756-1870 MDS - Final 2D Configuration - Scatterplot and Shepard Diagram	30
Figure 14	Wheat Price for Nine Regions - Period : 1756-1870 MDS - Final 2D Configuration - Multiple Scatterplot	31

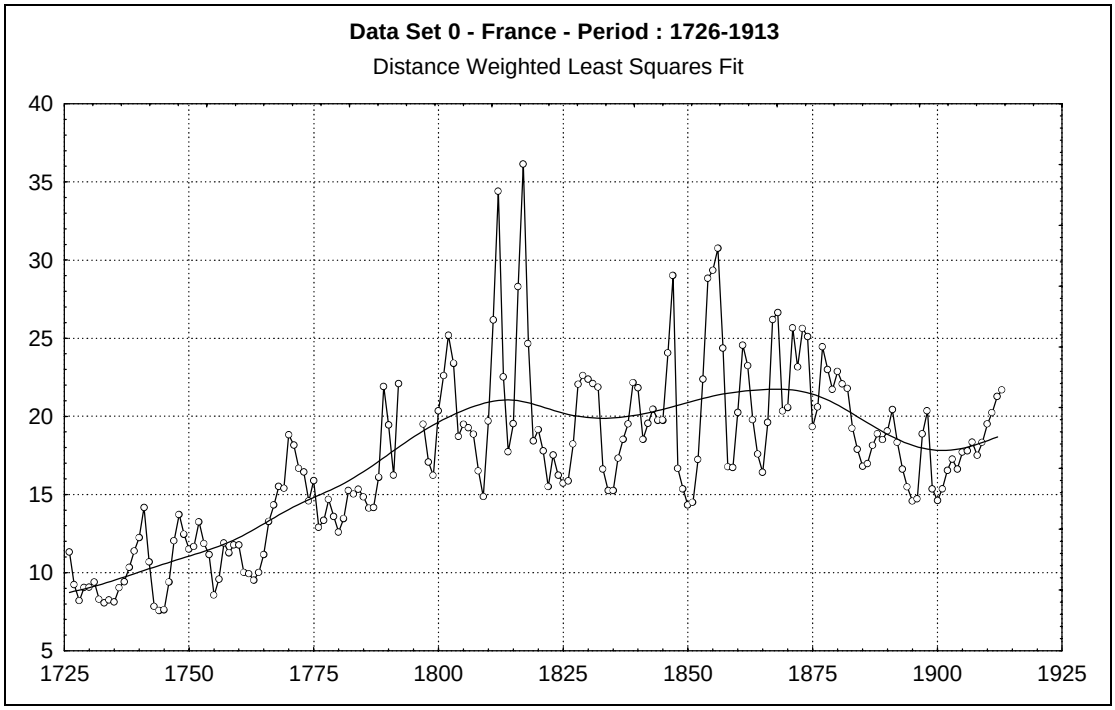
Section 1 : Introduction

The objective of this paper is to investigate the evolution of regional price differences of wheat in France during the period 1750-1870. More in particular, attention will be paid to the changing regional differences of this price evolution of grain.

The period of investigation is of particular interest. This period will be split up into three sub-periods. The first of these sub-periods is the period 1750-1790, in other words the period starting with the Seven Years War (1756-1763) and ending with the French Revolution in 1789. The results and conclusions for these four decades, which are in fact the last period of the Ancien Régime, will then be compared with the post-revolution period. For reason of convenience this post-revolutionary period will be split up into the sub-periods 1790-1830 and 1831-1870.

In the first place the available data consists of a general aggregated price series about the price of wheat in France for the period 1726-1913. For more precise information about this price series see Appendix 1. A graphical representation of this price series is given in Figure 1. The representation of this aggregated series is overlaid with the distance weighted least squares fit. This fitting curve is a nonlinear robust trend fitting procedure that can be compared with a weighted moving average of the observed values.

Figure 1 : General Aggregated Wheat Price for France - Period : 1726-1913
Sequential Line Plot - Distance Weighted Least Squares Fit



The graphical representation of the general aggregated price series enables to localize the period under investigation. The first sub-period 1750-1790 can be characterized by a substantial price increase. After the turbulent period caused by the French revolution the price variable fluctuated, until 1875, around a slightly increasing mean. A closer look at the evolution of the general price variable reveals that in 1870 the mean of this price level was roughly comparable with the price level at the beginning of the century. Apart from this relative stability in mean value, the period 1800-1870 was characterized by four pronounced peak values, i.e. the years 1812, 1817, 1847 and 1856. However the distance weighted least squares fit seems to be hardly affected by these peak values.

In this paper the breakdown of the aggregated price series into regional series will be based on nine regional series. These nine disaggregated time series are representing the price behavior of wheat in nine geographical areas called *'Grands Secteurs Teritoriaux'*. They are presented in Table 1. More

precise information about the exact composition of these regions could not be traced. The only information that could be found is from **Labrousse et al.** [**11**, p. 21], i.e. '*Secteurs constitués au XIXe siècle par les services nationaux de la statistique agricole. On s'est efforcé de grouper les généralités de l'ancien régime dans le cadre de ces secteurs'*. Further references about these data can be found in Appendix 1.

Table 1 : Nine Regions - Grands Secteurs Teritoriaux

8	**1**	**2**
North-West	**North**	**North-East**
7	**9**	**3**
West	**Center**	**East**
6	**5**	**4**
South-West	**South**	**South-East**

The nine regional series are graphically presented in Figure 2. The observed series are plotted in the left column, while the distance weighted least squares fit is given in the right column. In the first row the behavior of the nine series is presented for the whole period 1756-1870. From row two to four the regional price variables are presented for the sub-periods 1756-1790, 1797-1830 and 1831-1870 respectively.

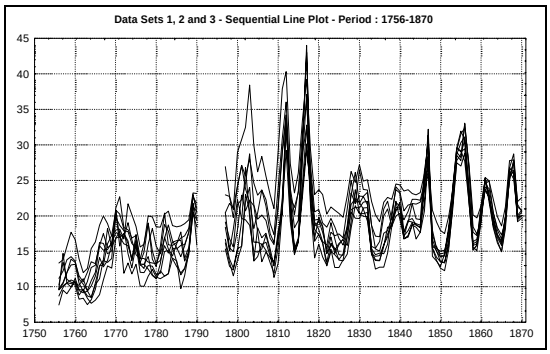
The characteristics that could be found in the aggregated price series can also be detected in the regional price variables. Apart from these general remarks the nine regional series reveal valuable and more detailed information about the price behavior of wheat. The additional information can be summarized in the following conclusions.

Conclusions

- All regions are responding to the peak years. This response is more pronounced in the third sub-period than in the second.
- Regional prices prior to 1810 are behaving asynchronously. Even the more robust distance weighted least squares fit is supporting this rather chaotic behavior.
- For the periods 1810-1820 and 1845-1870 the price evolution of the nine regions was more synchronous. This characteristic is emphasized by the distance weighted least squares fit.
- The period 1815-1855 is characterized by a general and gradually decline of the difference between the price levels of the nine regions.
- From 1855 on the difference in price level between the regional price variables seems to remain fairly constant.

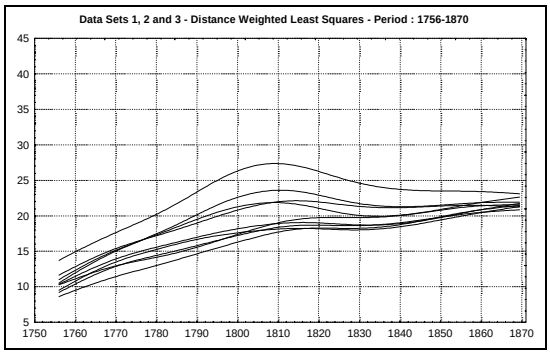
Figure 2 : Wheat Price for Nine Regions - Period : 1756-1870
Sequential Line Plot - Distance Weighted Least Squares Fit

Sequential Line Plot



Period : 1756-1870

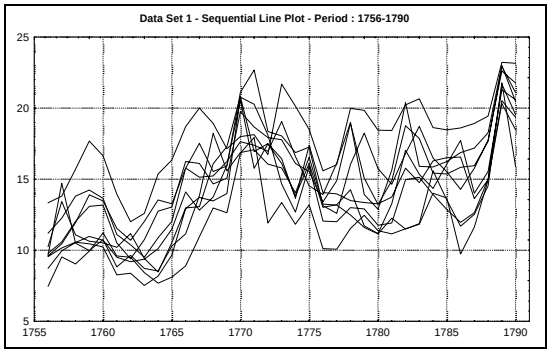
Distance Weighted Least Squares Fit



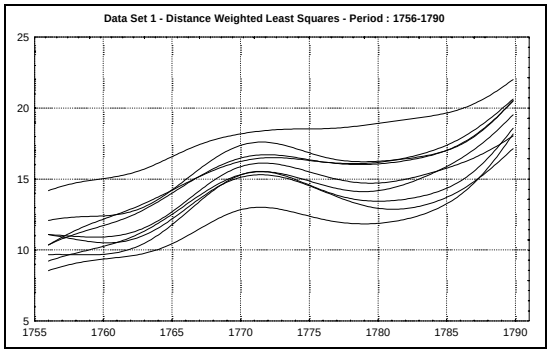
Period : 1756-1870

Figure 2 : Wheat Price for Nine Regions - Period : 1756-1870
Sequential Line Plot - Distance Weighted Least Squares Fit (Continued)

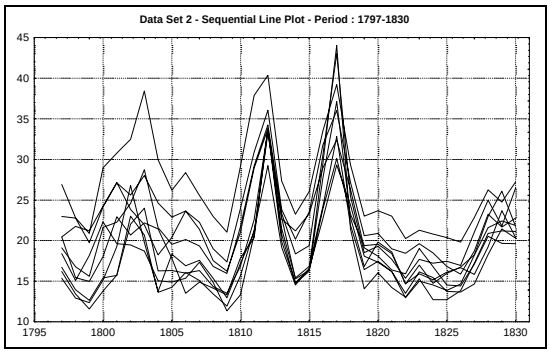
Sequential Line Plot | **Distance Weighted Least Squares Fit**



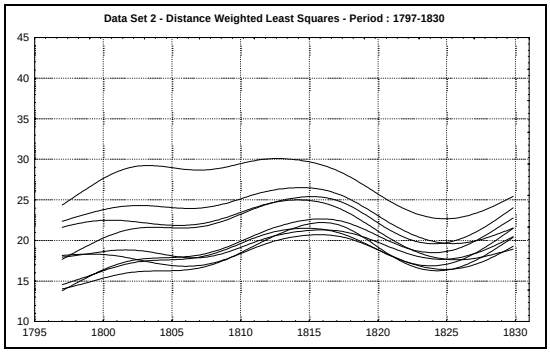
Period : 1756-1790 - Data Set 1



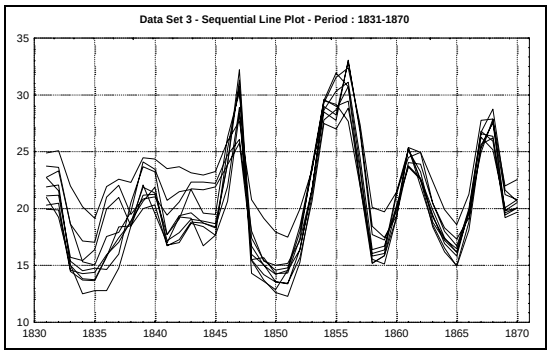
Period : 1756-1790 - Data Set 1



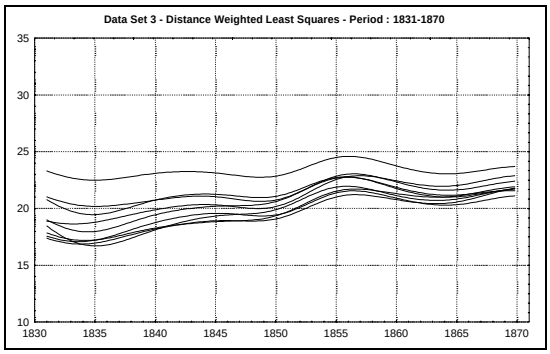
Period : 1797-1830 - Data Set 2



Period : 1797-1830 - Data Set 2



Period : 1831-1870 - Data Set 3



Period : 1831-1870 - Data Set 3

Section 2 : Box-and-Whisker Plot

Introduction

A box-and-whisker plot provides a simple graphical display of the distribution of a single variable. It contains much of the same information contained by five-number-summaries and letter-value displays, but summarizes the information graphically. Each box plot consists mainly of three elements. A first element is the box, including the values of the variable situated between the hinges, i.e. between the first and the third quartile. It follows that the width of the box is equal to the interquartile distance and that the box contains 50% of the data. The median or second quartile is marked by a '●' sign.

A second element of a box plot consists of the left and right whisker. These whiskers start at the hinges, i.e. the first or third quartile, and end at 1.5 times the interquartile distance from these hinges. Data values falling outside the whiskers are marked separately. These individual points are the third part of a box plot. Data values situated between 1.5 and three times the interquartile distance from the hinges are called outliers and are marked by a '0' symbol. Data values situated even more than three times the interquartile distance from the hinges are called extreme values and will be marked by an asterisk.

Results

The comparison of the box-and-whisker plots for the nine regions can best be based on the parallel box-and-whisker plots. These parallel plots are box-and-whisker-plots drawn side by side, oriented in the same direction, drawn to the same scale and lined up with the same axis. These parallel box plots provide a useful visual way to display and contrast the distributions of the multiple price series. The parallel box plots for each of the three periods are presented in Figure 3. In order to facilitate the comparison of the medians the median values are interconnected with each other.

The comparison of the different box plots will be mainly based on the following four characteristics :

- position of the median
- symmetry between the first and third quartile
- symmetry of values below and above the hinges
- outliers and extreme values

Conclusions

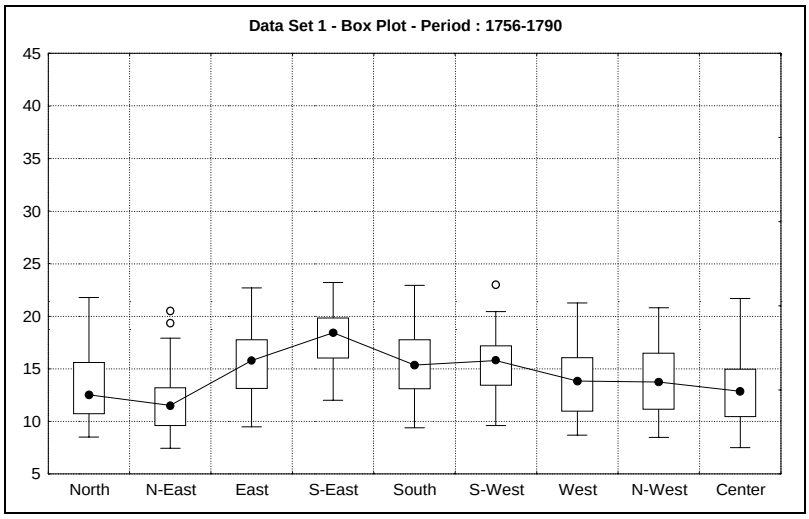
The comparison of the median values within each of the three periods reveals that for each of these periods the median values for the regions East, South-East, South and to a lesser degree South-West are definitely larger than for the other regions. The largest differences are obtained for the period 1797-1830 while for the period 1831-1870 these differences, although still present, are not so pronounced anymore. One can conclude that, for each of the three periods and using the median as a robust measure for central tendency, the price level of wheat was higher in the eastern region and the three southern regions than it was in the other six regions.

Whereas the asymmetry within the box, i.e. the asymmetry around the second quartile or median, seems to be of minor importance, this is not the case for the asymmetry of the data values beyond the two hinges. For the period 1797-1830 the wheat prices for all the regions, with the exception for the regions north-west and center, are characterized by a right-skewed distribution. In other words, the distribution of the data stretches to the right (higher values) more than it does to the left (lower values). Also for the period 1831-1870 the data values are not equally distributed. Also for this period prices for all regions, including the regions north-west and center, are positively skewed, i.e. the values are tailing off towards the high end of the scale. A tentative explanation for this characteristic might be the presence of the peak values around the years 1812 and 1817 and the quasi cyclic behavior of the price series in the period 1831-1870.

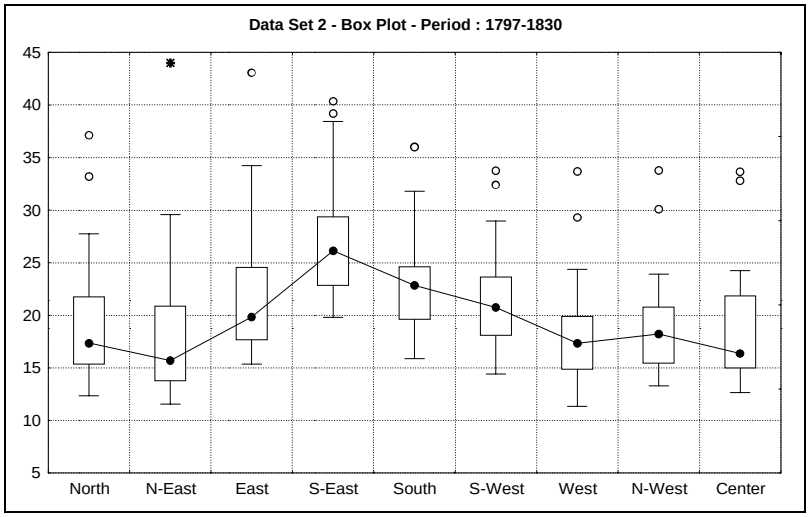
The influence of the peak values observed around the years 1812 and 1817 does not seem limited to the skewness of the distributions of the data. These data points resulted in two outliers for each of the

regions. For the region north-east this influence was so pronounced that it even resulted in a couple of extreme values.

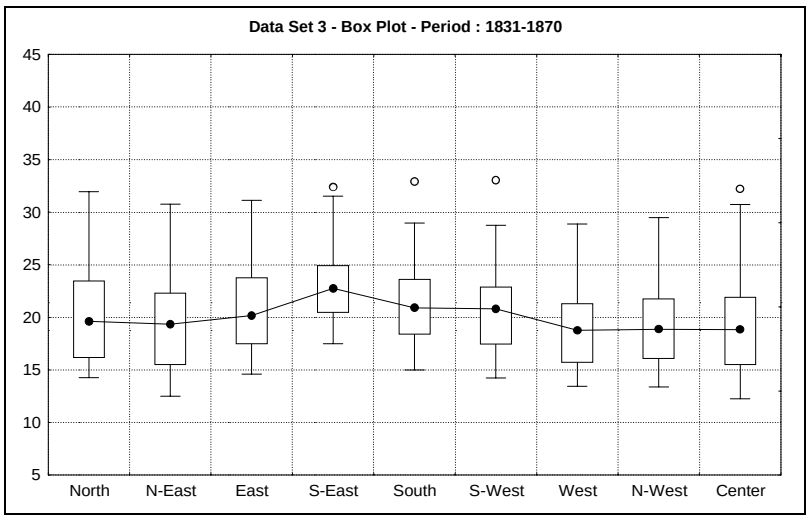
Figure 3 : Wheat Price for Nine Regions - Period : 1756-1870
Multiple Box-and-Whisker Plots - Connected Medians



Period : 1756-1790 - Data Set 1



Period : 1797-1830 - Data Set 2



Period : 1831-1870 - Data Set 3

A remarkable result contrasting to the influence of the peak years 1812 and 1817 is that the peak values in the period 1831-1870 didn't result in extreme values or even outliers for each of the series. It seems that the peak years 1839, 1847, 1856, 1861 and 1868 only affected the central and the southern regions. This could be an indication that these observations could be considered as part of a more general quasi cyclical behavior and not as isolated exceptional observations.

Section 3 : Correlation Analysis

Introduction

In this section we will investigate the association between the price series for the regions. Not only the existence of such an association will be investigated but also the magnitude of this association will be analyzed. Apart from the comparison of the different regions with each other also the comparison of the three periods will be investigated in more detail.

An obvious choice among the statistical measures of association is the product-moment or **Pearson** correlation coefficient. This correlation coefficient provides a symmetric measure of association between two variables both measured at the interval level. A linear relationship between the variables is assumed. It is perhaps the most used and misused measure of association in statistics. Therefore, it was decided that prior to its effective use in the analysis the necessary conditions had to be checked.

A preliminary and necessary condition in order to use the **Pearson** correlation coefficient is that both variables involved must be measured at the interval level. Since the data used consist of price series this condition is definitely fulfilled. A second condition is that for a meaningful use of this measure of association the relationship must be linear. In order to get an idea of the linearity of the relationship between the series the scatterplot of the series can be overlaid with a polynomial of higher order or with the more robust distance weighted least squares fit. This last procedure can be seen as a more robust and adaptive measure for the fundamental and underlying trend. For each of the three periods the scatterplots are given in Figure 4. Since almost none of the bivariate relationships reveals a clear-cut nonlinear relationship, it was decided to use the simple correlation coefficient as a measure of association.

Apart from all these considerations it must be mentioned that the most important advantage of using the **Pearson** correlation is the characteristic that this measure of association is invariant for a linear transformation. This means that it is perfectly allowed that the variables are expressed in different monetary (price) units and/or different measures for volume or weight.

Results

A real and practical problem to perform a detailed analysis of the association between the price series is the total number of correlation coefficients involved in the analysis. With nine regions and three periods this total number of coefficients amounts to a total of $3 \times 9 \times 9 = 243$ coefficients. Even if one takes the diagonal elements (the correlation of a variable with itself is equal to one) and the symmetry of the correlation matrices (the correlation between variable a and b is equal to the correlation between variable b and a) into account the total number of coefficients for interpretation is still equal to $3 \times 36 = 108$ coefficients. These three correlation matrices are tabulated in Table 2.

The interpretation of these coefficients can be considerably simplified by using an appropriate graphical representation. An often used representation takes the form of a matrix plot. The off-diagonal elements of this matrix plot consist of the individual scatterplots each of which can be overlaid by one or other fitting curve. On the main diagonal an histogram represents the distribution of the variables. For each of the three periods the matrix plot is given in Figure 4. The fitting curve used with these matrix plots is the distance weighted least squares fit.

For this specific and particular application two alternative graphical representations can be constructed. In a first graph the correlation coefficients between the central and the outer regions will be represented by using a circular graph. Given the geographical meaning of the nine regions it is obvious to situate the central region in the center of the circle and the outer regions on the circumference of the circle. The outer regions are equally spaced and ordered clockwise, starting with the region north at the top of the graph. The correlation coefficients are plot on the radii and are scaled between 0.5 and one. To facilitate the interpretation of the correlations, five concentric circles are used as scale lines. For each of the three periods the marked correlations are connected. This graph is presented in Figure 5.

Table 2 : Wheat Price for Nine Regions - Period : 1756-1870
Correlation Matrices

	Period : 1756-1790 - Data Set 1								
	North	**N-East**	**East**	**S-East**	**South**	**S-West**	**West**	**N-West**	**Center**
North	1.00	0.82	0.76	0.57	0.70	0.64	0.75	0.87	0.84
N-East	0.82	1.00	0.88	0.69	0.80	0.74	0.80	0.82	0.89
East	0.76	0.88	1.00	0.82	0.90	0.77	0.83	0.81	0.95
S-East	0.57	0.69	0.82	1.00	0.87	0.84	0.74	0.63	0.70
South	0.70	0.80	0.90	0.87	1.00	0.95	0.91	0.79	0.90
S-West	0.64	0.74	0.77	0.84	0.95	1.00	0.88	0.74	0.78
West	0.75	0.80	0.83	0.74	0.91	0.88	1.00	0.93	0.90
N-West	0.87	0.82	0.81	0.63	0.79	0.74	0.93	1.00	0.90
Center	0.84	0.89	0.95	0.70	0.90	0.78	0.90	0.90	1.00

	Period : 1797-1830 - Data Set 2								
	North	**N-East**	**East**	**S-East**	**South**	**S-West**	**West**	**N-West**	**Center**
North	1.00	0.94	0.88	0.70	0.76	0.70	0.83	0.85	0.94
N-East	0.94	1.00	0.95	0.75	0.82	0.75	0.80	0.81	0.93
East	0.88	0.95	1.00	0.89	0.91	0.81	0.82	0.85	0.94
S-East	0.70	0.75	0.89	1.00	0.95	0.85	0.74	0.78	0.83
South	0.76	0.82	0.91	0.95	1.00	0.94	0.83	0.83	0.85
S-West	0.70	0.75	0.81	0.85	0.94	1.00	0.84	0.72	0.75
West	0.83	0.80	0.82	0.74	0.83	0.84	1.00	0.89	0.87
N-West	0.85	0.81	0.85	0.78	0.83	0.72	0.89	1.00	0.89
Center	0.94	0.93	0.94	0.83	0.85	0.75	0.87	0.89	1.00

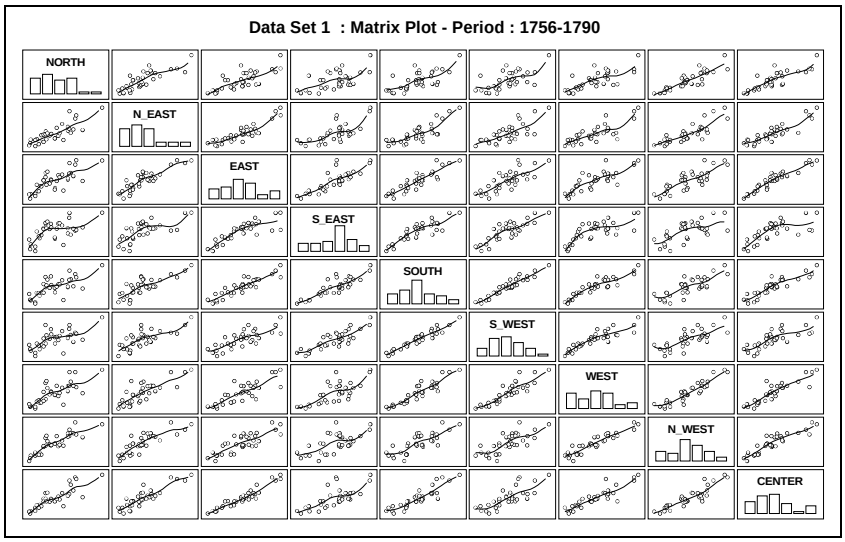
	Period : 1831-1870 - Data Set 3								
	North	**N-East**	**East**	**S-East**	**South**	**S-West**	**West**	**N-West**	**Center**
North	1.00	0.98	0.97	0.91	0.87	0.89	0.96	0.97	0.97
N-East	0.98	1.00	0.96	0.89	0.87	0.89	0.96	0.95	0.96
East	0.97	0.96	1.00	0.96	0.91	0.90	0.94	0.93	0.97
S-East	0.91	0.89	0.96	1.00	0.97	0.94	0.92	0.90	0.93
South	0.87	0.87	0.91	0.97	1.00	0.98	0.93	0.89	0.93
S-West	0.89	0.89	0.90	0.94	0.98	1.00	0.96	0.92	0.94
West	0.96	0.96	0.94	0.92	0.93	0.96	1.00	0.98	0.99
N-West	0.97	0.95	0.93	0.90	0.89	0.92	0.98	1.00	0.97
Center	0.97	0.96	0.97	0.93	0.93	0.94	0.99	0.97	1.00

An alternative for using a matrix plot to represent the correlations among the outer regions is to use a separate plot for each of the eight outer regions. In other words, we will use a separate plot for each row or column of the reduced matrix plot, i.e. excluding the row and column for the central region. Each of these eight graphs is constructed following exactly the same design. The value of the correlations is on the vertical axis, scaled between 0.5 and one, and the eight outer regions, ordered clockwise, on the horizontal axis. Each graph starts with another region and represents the correlations between this first region and the other outer regions. For each graph the sequence of outer regions is appended with the starting region on the ninth position. In other words the first and last region in the sequence will be the same. This has the advantage that the region in the middle of this sequence will be diagonally opposed to the first and the last outer region on the graph. The resulting eight graphs can be found in Figure 6.

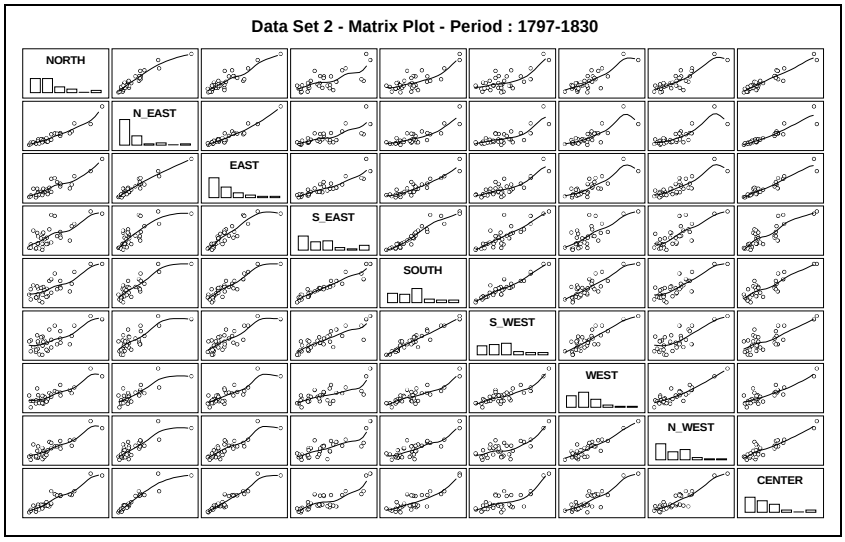
Conclusions

A first conclusion is that all the correlation coefficients have a positive sign. Since the correlation coefficient is a standardized symmetric measure of association with values that can range from -1 to +1, this means that all price series are characterized by a positive association. In other words, this means that for all series the observations do have the tendency to be situated on the same side of the mean value of the series. Another interpretation of these positive correlations might be that all series are predominantly evolving in the same direction.

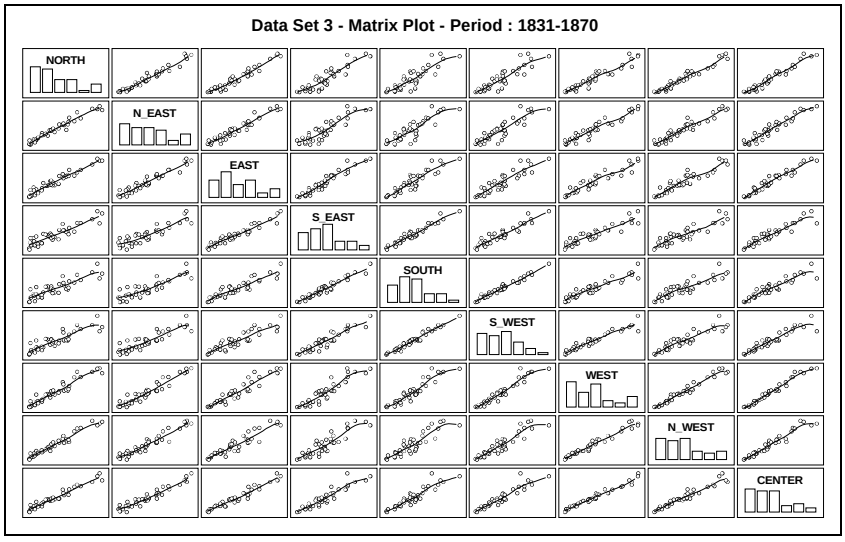
Figure 4 : Wheat Price for Nine Regions - Period : 1756-1870
Matrix Plots - Histogram - Scatterplot - Distance Weighted Least Squares Fit



Period : 1756-1790 - Data Set 1

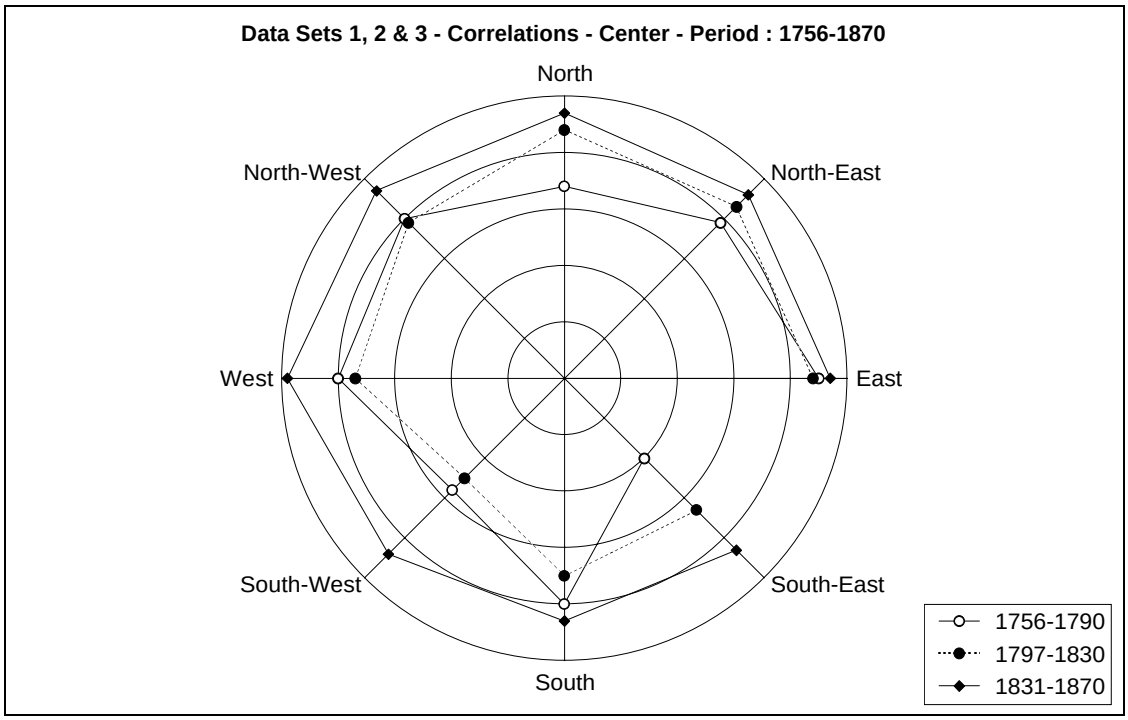


Period : 1797-1830 - Data Set 2



Period : 1831-1870 - Data Set 3

Figure 5 : Wheat Price for Nine Regions - Period : 1756-1870
Correlations between Center and Outer Regions

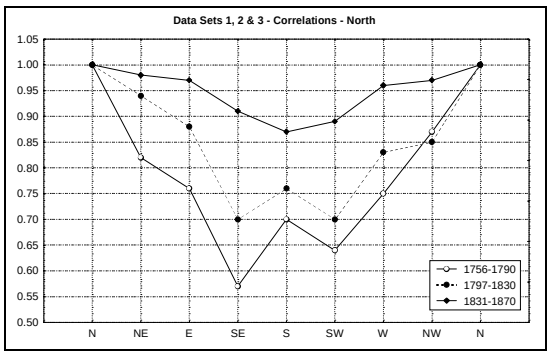


A second conclusion that can be drawn is about the association between the central and the outer regions. From the graphical representation in Figure 5 it can be seen that the correlations for the period 1797-1830 are rather close to those for the period 1756-1790. The main exceptions are the correlations between the central region and the regions north and south-east. For these correlations there seems to be a substantial increase of both correlation coefficients. A last conclusion about the association between the center and the outer regions is that without any exception the correlations for the period 1831-1870 are higher than those for the two previous periods. Besides it turns out that all these correlations are situated between the values 0.9 and 1.0.

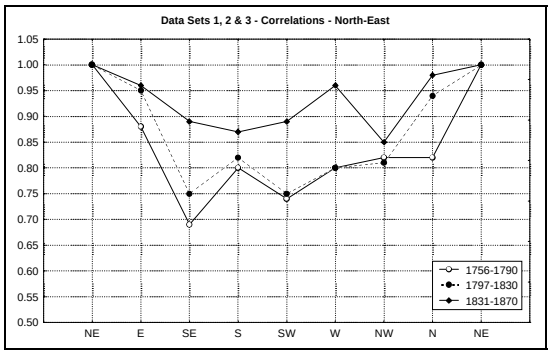
A third conclusion is about the intercorrelations between the outer regions. For all the outer regions the correlations with neighboring regions are higher than with the more remote regions. It can be seen from Figure 6 that for almost all the regions the correlation has the tendency to decrease with more distant regions. This last conclusion is based on the U-shaped representation of the correlation coefficients in Figure 6. Furthermore, the visual inspection of this graphical representation reveals that the largest correlations between the outer regions are obtained for the period 1831-1870.

The general conclusion is then that for both the correlations between the outer regions and the central region and the intercorrelations between the outer regions the results for the periods 1756-1790 and 1797-1830 are roughly comparable, while the correlations obtained for the period 1831-1870 are substantially higher.

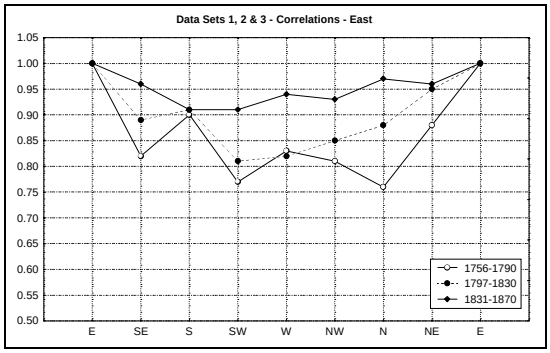
Figure 6 : Wheat Price for Nine Regions - Period : 1756-1870
Correlations among Outer Regions



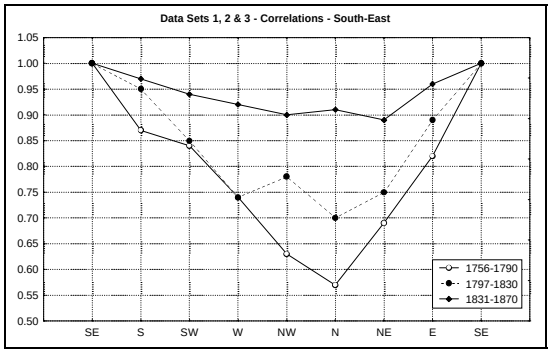
North



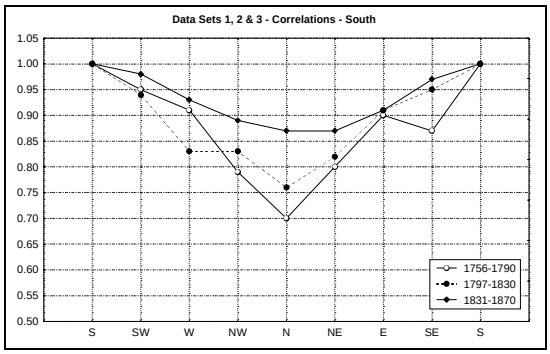
North-East



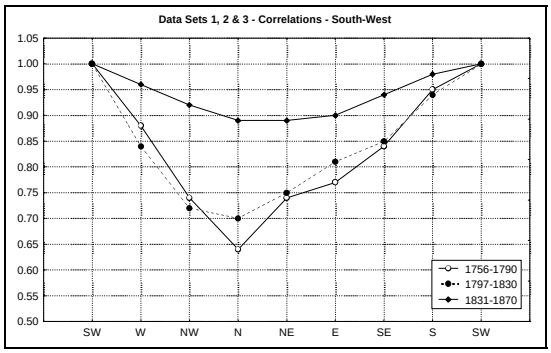
East



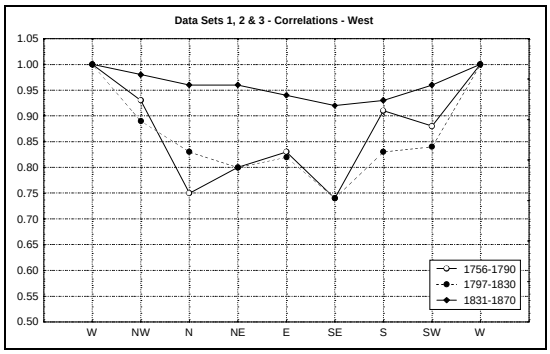
South-East



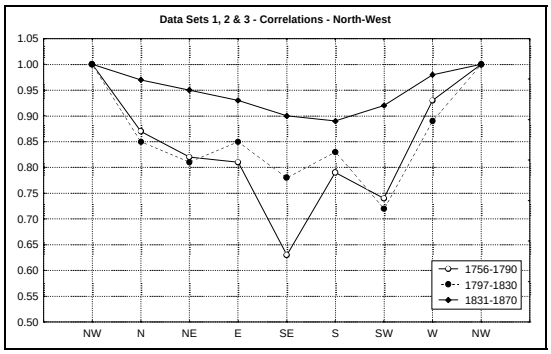
South



South-West



West



North-West

Section 4 : Principal Component Analysis - Component Loadings

Introduction

Principal components analysis is a multivariate statistical technique that linearly and systematically transforms an original set of correlated variables into a substantially smaller set of uncorrelated variables that still represents most of the information in the original set of variables. The final goal is to reduce the dimensionality of the original data set in the hope that a smaller set of uncorrelated variables will be much easier to understand than a larger set of correlated variables. In other words the goal of this technique is trying to explain a maximum of the variation in the original observed data on the basis of a few underlying dimensions.

This new and smaller set consists of unobserved and uncorrelated variables, called (principal) components or factors. These components are common to all of the observed variables and may be interpreted as a set of fundamental, underlying (latent) variables. These hypothetical constructs can also be seen as a set of common uncorrelated explanatory variables. Each of the observed variables can then be written as a linear combination of principal components, acting as explanatory variables.

It must be noted that in the first place principal components analysis is a deterministic and mathematical technique which does not require to specify an underlying statistical model to explain the error structure of the data. No particular assumption needs to be made about the probability distribution of the original variables. It is a variable-directed technique which is appropriate when the variables arise 'on equal footing' so that no distinction needs to be made between dependent and explanatory variables. Principal component analysis is also an exploratory technique and should be used to get the 'feel' of the data. Hopefully, the method will lead to a better understanding of the correlation structure and may generate hypotheses regarding the relationship between the variables.

The principal component analysis typically searches for a set of uncorrelated linear combinations of the original variables that captures most of the information in these original variables. These linear composites are constructed in such a way that the first of these components explains the maximum of the total variability in the data and the remaining components will, each in turn, explain the maximum of the remaining variance in the observed data. In this way the final solution will consist of the principal components listed in the order of the percent of variance for which they account.

The problem remains how many components ought to be retained in the analysis. Among the many criteria that have been suggested the most important are the **Kaiser** and the **Cattell** criterion. Both these criteria have been studied in detail. The final conclusion from these comparisons is that Kaiser's criterion typically results in too many included principal components, while Cattell's criterion typically includes too few. However, both criteria do quite well under normal conditions, that is to say when there are relatively few components and many variables. (See **Hakstian et al. [6]** and **Zwick & Velicer [10]**)

Cattell's criterion is based on the scree test, i.e. a graphical method that consists in the interpretation of a scree plot. In a scree plot the eigenvalues of the correlation matrix, representing the variances of the principal components, are arranged in descending order and plotted against their indices. Generally this plot breaks visually into a steady downward slope (mountain side) and a gradual tailing away (scree side). The break from the steady downward slope indicates the break between the important principal components and the remaining components which make up the scree.

The principal component analysis need not to be restricted to just a data reduction method. This technique can also be applied to detect structure in the relationships between variables. In this respect the principal component technique can also be used as a classification method. Using this technique as a classification method will be based on the component loadings. These component loadings are the correlation coefficients between the variables and the components. They provide a convenient summary of the influence of the components on the variables and thus a useful basis for information about these components. It is exactly the pattern of these loadings that will be helpful in 'labeling' the hypothetical constructed components.

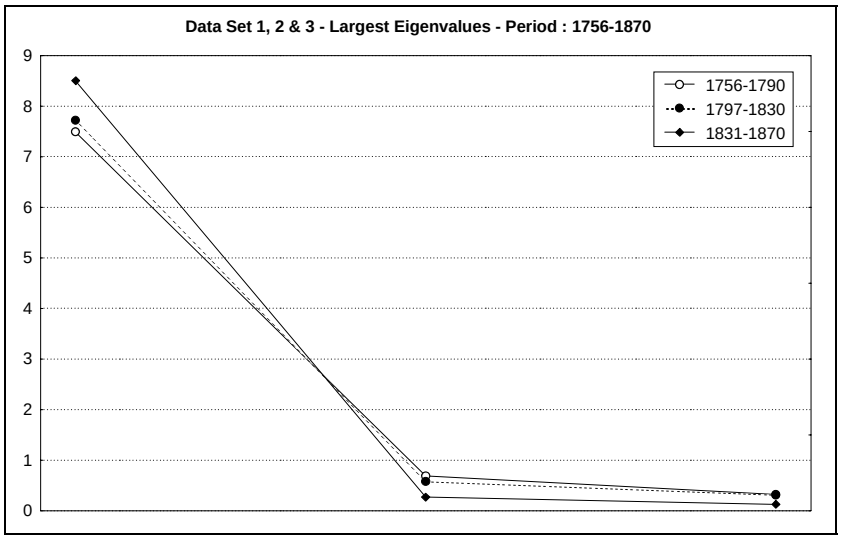
However, these component loadings are inherently indeterminate. They are uniquely determined up to an orthogonal transformation or rotation. Consequently, any solution can be rotated arbitrarily to arrive at a new solution. In practice this indeterminacy is used to arrive at a component solution that has what is called a 'simple structure'. A component solution is said to have a simple structure if each variable is loaded highly on one component and all component loadings are either large in absolute value or near zero otherwise.

Many different objective criteria have been developed for choosing the appropriate rotation procedure. For this application the varimax rotation was chosen. The varimax rotation of the component loadings is aimed at maximizing the variances of the squared raw component loadings across variables for each component. This rotation procedure is equivalent to maximizing the variances in the columns of the matrix of component loadings.

Results

A first step in performing a principal component analysis is the decision about the number of components that should be retained. This decision will be based on the scree test. A visual representation of this test is given in Figure 7. For each of the three sub-periods this figure represents the three largest eigenvalues of the correlation matrices of the nine regional price series. Using this scree plot it was decided that for each of the three periods two components had to be retained for further analysis. Regarding this scree plot the close resemblance between the results for the periods 1756-1790 and 1797-1830 must be mentioned. Especially for the period 1831-1870 the arguments in favor of a two-component model are even more pronounced.

Figure 7 : Wheat Price for Nine Regions - Period : 1756-1870
Largest Eigenvalues - Scree Plot



Using just two components for each of the three periods means that one needs only two exploratory variables to explain or approximate the variability or variance of the nine price variables. These two common principal components account for respectively 90.8%, 92.2% and 97.5% of the total variance of the price variables. The unexplained proportion of the variances, ranging from 2.5% to almost 10%, must be considered as lost information and may not be interpreted as specific and variable related variance.

Detailed and numerical results for these two-component solutions are tabulated in Table 3. For each of the three periods 1756-1790, 1797-1830 and 1831-1870 this table reports the component loadings as well as the total variance explained by each of the two common explanatory components. The results were obtained after the rotation of the initial and original principal component solution by means of a varimax rotation.

Table 3 : Wheat Price for Nine Regions - Period : 1756-1870
Two-Component Model - Component Loadings - Varimax Rotation

	1756-1790		1797-1830		1831-1870	
Variables	**Comp. 1**	**Comp. 2**	**Comp. 1**	**Comp. 2**	**Comp. 1**	**Comp. 2**
North	0.900530	0.282854	0.910776	0.359602	0.853657	0.511624
North-East	0.784442	0.493229	0.844464	0.455830	0.845651	0.509052
East	0.669496	0.663402	0.742919	0.620473	0.760044	0.616906
South-East	0.291537	0.899694	0.443954	0.847023	0.584522	0.784593
South	0.516410	0.840176	0.498034	0.859417	0.505597	0.859405
South-West	0.418158	0.859994	0.390679	0.879134	0.558353	0.813218
West	0.671406	0.664935	0.692728	0.596892	0.744746	0.648761
North-West	0.856285	0.429998	0.763238	0.521622	0.807091	0.561117
Center	0.800743	0.550403	0.843276	0.500319	0.767131	0.627132
Explained Var.	**0.469584**	**0.438700**	**0.496205**	**0.425521**	**0.525012**	**0.449554**
Total	**0.908284**		**0.921726**		**0.974566**	

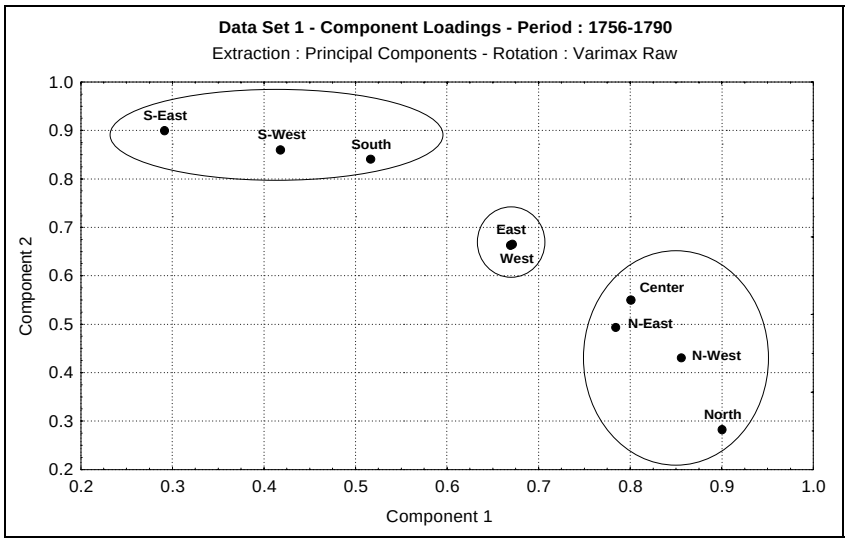
In Table 3 the component loadings larger than 0.7 are printed on a gray background. If the explanatory power of a component on a given variable is quantified by a component loading of 0.7, this means that this component accounts for almost 50 percent (the squared loading) of the variance of that variable. The component loadings after varimax rotation for the two-component case can also be presented graphically. In Figure 8 a scatter plot of the two components is given for each of the three periods.

Conclusions

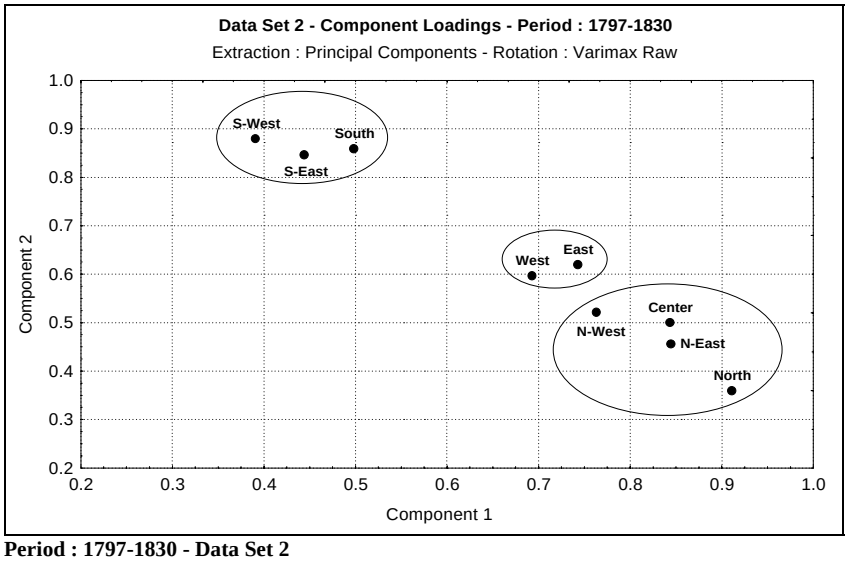
Combining the results of Table 3 with the graphical representation of Figure 8 leads to the following conclusions :

- the central and northern regions are highly correlated with the first component. This component can be seen as the exploratory variable for the price series in the center and the northern part of France. The four regions influenced by this first component are graphically represented by a first cluster.
- the highest loadings with the second component are those with the regions south-east, south and south-west. From this one can conclude that this second component acts as the exploratory variable for the price series in the southern part of France. Also these regions are represented by a second cluster.
- for the period 1756-1790 the regions east and west are equally explained by both the first and the second component. None of these components is dominating the other. With other words this means that the price series for both regions are behaving between the price series in the northern and the southern part of France. Both series are graphically represented by a separate but smaller cluster.
- during the period 1797-1830 this smaller cluster shifted to the cluster containing the southern regions. From the results for the period 1831-1870 one can see that the regions east and west have become an integral part of the larger cluster with the southern regions.
- the size of the two larger clusters is gradually becoming smaller. This means that for the northern part as well as for the southern part of France the price differences between regions within each of these larger clusters are becoming smaller.
- the distance between the two larger clusters is gradually becoming smaller. In other words not only within each of the larger clusters but also between these clusters the differences between the price series are becoming smaller.

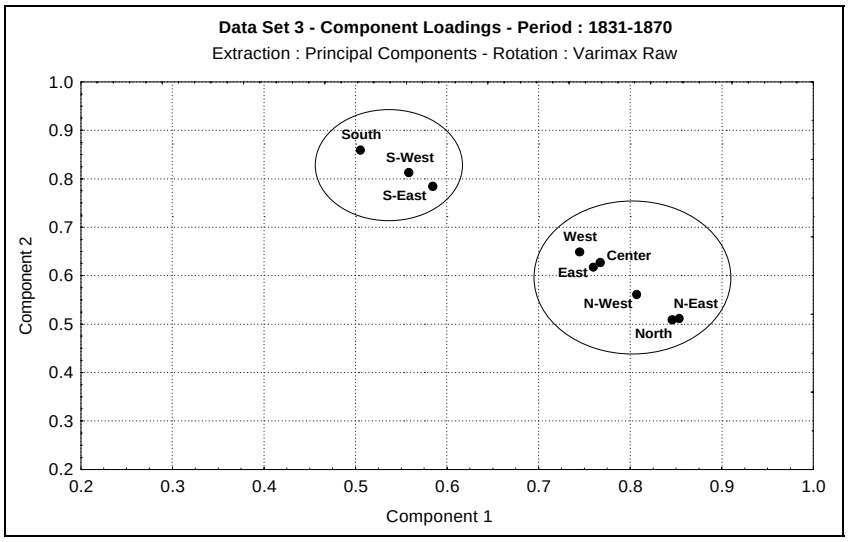
Figure 8 : Wheat Price for Nine Regions - Period : 1756-1870
Two-Component Model - Varimax Rotated Loadings - Scatter Plot



Period : 1756-1790 - Data Set 1



Period : 1797-1830 - Data Set 2



Period : 1831-1870 - Data Set 3

Section 5 : Cluster Analysis - Original Data

Introduction

Cluster analysis is a collection of different algorithms and has nothing to do with the typical statistical significance testing. It is used when one does not have any a priori hypotheses but is still in the exploratory phase of the research. The basic idea and aim of cluster analysis is to find the natural groupings, if any, of individual elements or variables. More formally, cluster analysis aims to allocate a set of individuals to a set of mutually exclusive and exhaustive groups such that individuals within a group are similar to one another while individuals in different groups are dissimilar.

A set of groups is called a partition. The groups forming a partition may be subdivided into smaller sets or grouped into larger sets so that one eventually ends up with a complete hierarchical structure of the given set of individuals. This structure is called hierarchical tree and can best be represented diagrammatically by using a dendrogram.

A dendrogram is a graphical display which shows the elements which combine to form a cluster and the level at which they form that cluster. A dendrogram is a convenient graphical representation for displaying the hierarchical formation of these clusters. It is placed on its side with branches running horizontally across the plot. The horizontal axis displays the level of similarity or dissimilarity at which the clusters form. The names of the elements or variables are displayed on the left of the branches.

Tree Clustering - Hierarchical Agglomeration Methods

The basic idea and common characteristic of the algorithms belonging to this category is to join together elements into successively larger clusters using some measure of similarity or distance. The larger these clusters the more the members of the cluster will be dissimilar. A typical final result of this approach is the agglomeration tree.

The method that will be used here is the method of **Ward [9]**. This method, which is known in literature under different names, is regarded as very efficient and tends to create clusters of small but comparable size.

k-Means Clustering

In general the k-means clustering method will produce exactly k different clusters of greatest possible distinction. Computationally the algorithm will start with k random clusters and then will move objects around from cluster to cluster with the final goal of minimizing the variability within the clusters and maximizing the variability between the clusters.

The results of a k-means clustering analysis consist of :

- the elements classified in each of the k clusters
- the distances between all the elements of the cluster and the center of the cluster to which they belong
- the distances between the k cluster centers

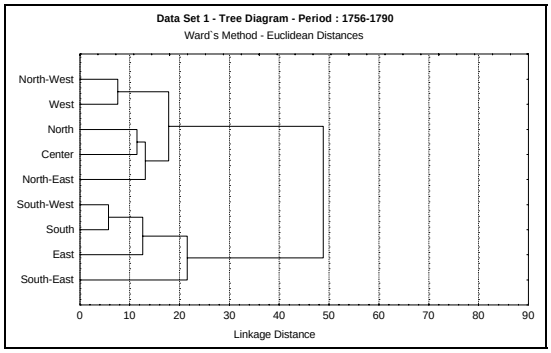
Another very useful additional result of the k-means clustering method is that for each cluster a representative variable for that cluster can be constructed. Each observation of this variable consists of the arithmetic mean of the corresponding observations of all the variables belonging to that cluster.

Results : Tree Clustering

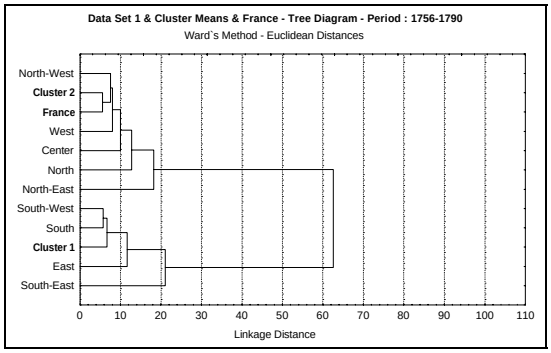
Ward's method was applied to each of the three data sets. The exact structure of the resulting hierarchical trees is presented by the three dendrograms displayed in Figure 9. Apart from these dendrograms detailed information about the composition of these clusters is tabulated in Table 4.

Figure 9 : Wheat Price for Nine Regions - Period : 1756-1870
Tree Clustering - Ward's Method - Structure of Clusters

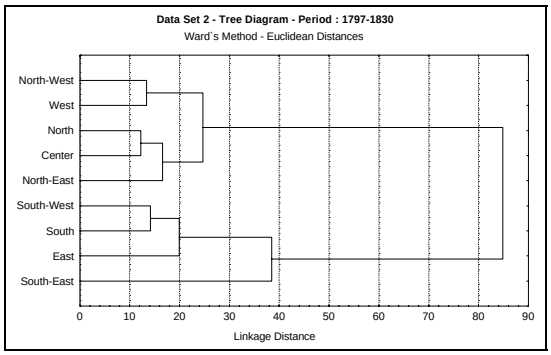
Original Data Sets | **Supplemented Data Sets**



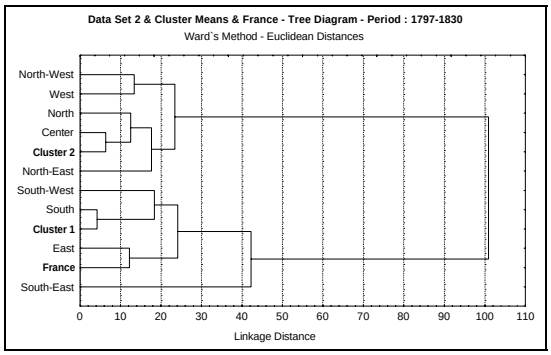
Period : 1756-1790 - Data Set 1



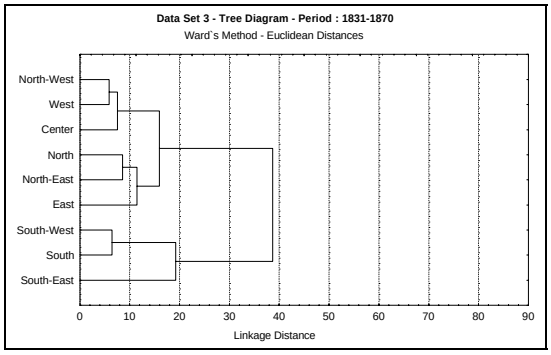
Period : 1756-1790 - Data Set 1



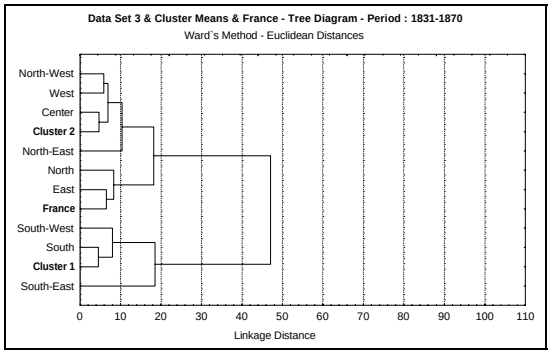
Period : 1797-1830 - Data Set 2



Period : 1797-1830 - Data Set 2



Period : 1831-1870 - Data Set 3



Period : 1831-1870 - Data Set 3

Table 4 : Wheat Price for Nine Regions - Period : 1756-1870
Tree Clustering - Ward's Method - Composition of Clusters

	Period											
	1756-1790				1797-1830				1831-1870			
	Clusters				Clusters				Clusters			
Region	2.1	2.2	1.1	1.2	2.1	2.2	1.1	1.2	2.1	2.2	1.1	1.2
Center		●				●			●	←		
North		●				●				●		
North-East		●				●				●		
East			●				●			●	←	
South-East				●				●				●
South			●				●				●	
South-West			●				●				●	
West	●				●				●			
North-West	●				●				●			

A first important result illustrated by the dendrograms is that for each of the three data sets the nine regional price series are categorized in two main clusters. Furthermore, it appears that each of these two main clusters can be further decomposed into two even smaller sub-clusters. The interpretation of clusters on still lower levels in the hierarchy is highly unreliable for the distances between these smallest clusters is becoming really small.

Conclusions : Tree Clustering

The results for the tree clustering can be summarized as follows :

- clustering for the periods 1756-1790 and 1797-1830 seems to be identical. Both periods are characterized by a first main cluster (Cluster 1) containing the region east and the three southern regions. One of the two sub-clusters contains only one region, i.e. the region east. Also the second main cluster (Cluster 2) consists of two sub-clusters. The first of these sub-clusters (Cluster 2.1) contains the regions west and north-west while the central region and the regions north and north-east are forming the second sub-cluster (Cluster 2.2).
- basically the clustering for the period 1831-1870 is identical to the results for the first and second period. The only two changes are the positions of the central region and the region east. Whereas in the previous periods the central region was allocated to the sub-cluster 2.2 this region shifts to the sub-cluster 2.1, i.e. much closer to the regions west and north-west. The second change is the reallocation of the region east. This region becomes part of the sub-region 2.2, i.e. closer to the regions north and north-east. On the level of the main clusters the results for the period 1830-1870 illustrate the partition of France in a southern part, consisting of the regions south-east, south and south-west, and a northern part containing all the other regions.

Table 5 : Wheat Price for Nine Regions - Period : 1756-1870
Cluster Members and Distances from Cluster Means

		Period		
Cluster	**Region**	**1756-1790**	**1797-1830**	**1831-1870**
Cluster 1	**East**	1.3185	2.4858	1.3201
	South-East	2.0966	3.8848	1.7976
	South	0.9955	0.7317	0.7181
	South-West	1.1746	2.6580	1.2399
Cluster 2	**West**	1.3372	2.0172	0.8379
	North-West	1.2585	1.9067	0.7775
	North	1.2874	1.6645	1.0146
	North-East	1.8342	2.3649	0.9796
	Center	1.0114	1.0853	0.7494

Results : k-Means Clustering

Given the results for the tree clustering method, it was decided to apply the k-means clustering in the case of two clusters. The results of the k-means clustering method for these two clusters are tabulated in Table 5.

Conclusions : k-Means Clustering

From these tabulated results the following conclusions can be drawn :

- for each of the three periods the composition of the two clusters is identical. Moreover, these results are almost wholly compatible with the two clusters obtained by **Ward**'s method. The only difference is that for the period 1830-1870 the region east remains in the cluster describing the price behavior in the southern part of France.
- using the results from Table 5 about the distances between the cluster members and their cluster center, it follows that these distances were by far the largest for the period 1797-1830 and the smallest for the period 1831-1870.

- for the cluster describing the southern part of France the region south is the closest region to the center of the cluster. For the second cluster, covering the northern part of France, the central region is the closest to the cluster center. These results hold for each of the three periods.

Another result from applying the k-means clustering method are the hypothetical and constructed variables representing the behavior of the cluster centers. A graphical representation of the two representative variables for each of the three periods is displayed in Figure 10. The raw variables for the cluster centers can be found in the left column of this figure, while in the right column the distance weighted least squares fit for these representative variables is presented. The price behavior for these cluster centers leads to the following additional conclusions :

- a first conclusion that can be drawn is about the level of the representative variables and is mainly based on the distance weighted least squares fit for these variables. The difference in level between the variables representing the northern and the southern cluster remains fairly constant until 1810 à 1815, i.e. about in the middle of the second period. From then on the price difference between the northern and southern cluster is becoming smaller and smaller. An even more precise conclusion can be drawn from the results summarized in Table 6. From these results it follows that the largest Euclidean distance between the two cluster centers was obtained for the period 1797-1830, while the smallest difference occurred for the last sub-period 1831-1870.
- a second conclusion is about the pattern of the raw data for the cluster centers. Comparing this pattern for each of the three periods leads to the conclusion that the synchronization between the two representative variables is much more pronounced in the last period than in the previous sub-periods. In other words the tendency of the two variables to move in the same direction seems to be the most evident for the last period. This aspect of synchronization can be quantified by using the **Pearson** correlation coefficient. For each of the three sub-periods the correlation coefficient between the two representative variables is given in Table 6. From these results it follows that for the first and second period the two cluster centers are highly and almost equally correlated, i.e. correlations of 0.88 and 0.89 respectively, while the highest correlation coefficient was obtained for the period 1831-1870. For this sub-period the correlation between the two clusters was 0.95.

Table 6 : Wheat Price for Nine Regions - Period : 1756-1870
Euclidean Distances and Pearson Correlation between Cluster 1 and Cluster 2

	1756-1790	**1797-1830**	**1831-1870**
Distance	3.1628	5.4017	2.3943
Correlation	0.8791	0.8886	0.9547

Enlarged Data Sets

In the next paragraphs attention will be paid to the exact position within each of the clusters of the representative cluster variables. Apart from the cluster centers also the position of the general aggregated price variable will be investigated. Therefore, the three data sets, each consisting of the nine regional price series, were completed with the two cluster variables and the aggregated price variable.

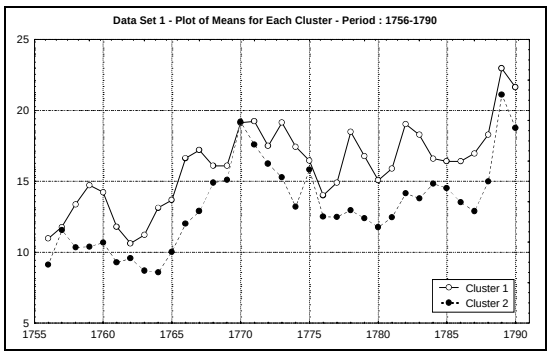
These supplemented data sets, each consisting of 12 series, were then used as input for a tree clustering analysis. The resulting dendrograms can be found as the right column of Figure 9. As could be expected, the composition of the resulting clusters for the enlarged data sets is exactly the same as for the original regional series. The only main difference is that the hierarchical structure of the clustering tree obtained for the second cluster of the first period has changed. From these results it can be concluded that for each of the three periods the representative variable for the first cluster is located close to the price variable for the region south. In contrast with the results for the first cluster, it follows that the location of the variable describing the price behavior for the second cluster center is shifting from the region north-west to the central region for the periods 1797-1830 and 1831-1870.

Some interesting remarks can also be made about the behavior, within the context of a clustering environment, of the general aggregated price variable for France. To better understand the correct location of this price variable, use was made of an additional graphical representation. This representation is displayed in Figure 11. For each of the three periods the distance weighted least

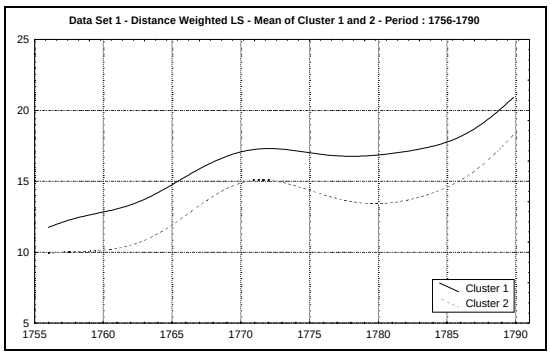
squares fit for the representative variables of the two clusters and the aggregated price variable are displayed in the right column. The nine regional variables are given in the left column of this graph.

Figure 10 : Wheat Price for Nine Regions - Period : 1756-1870
Cluster Means - Sequential Line Plot - Distance Weighted Least Squares Fit

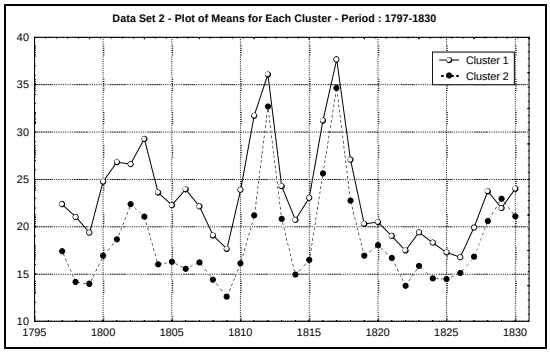
Sequential Line Plot | **Distance Weighted Least Squares Fit**



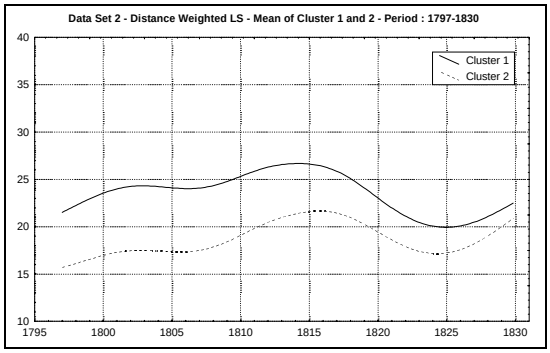
Period : 1756-1790 - Data Set 1



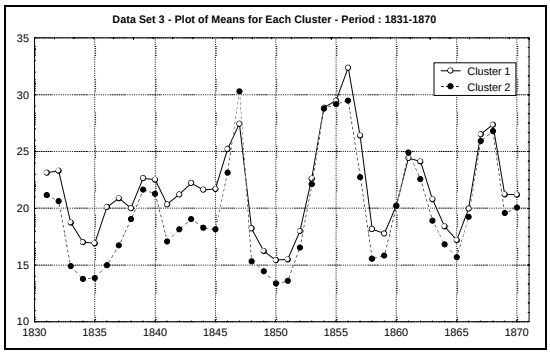
Period : 1756-1790 - Data Set 1



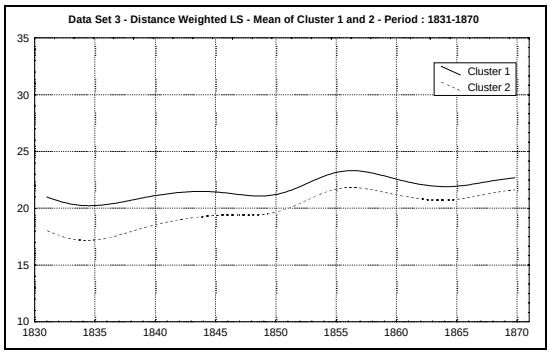
Period : 1797-1830 - Data Set 2



Period : 1797-1830 - Data Set 2



Period : 1831-1870 - Data Set 3



Period : 1831-1870 - Data Set 3

Conclusions : Enlarged Data Sets

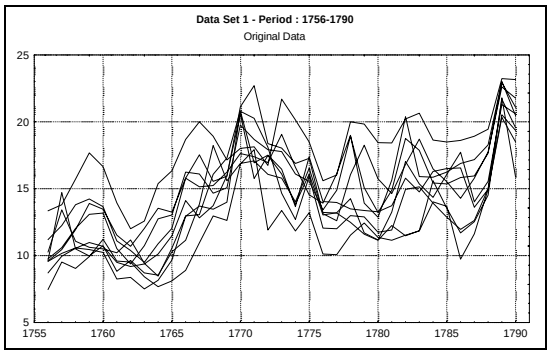
Combining the results presented in the right columns of Figure 10 and Figure 11 the following main conclusions can be drawn :

- for each of the three sub-periods the general aggregated price variable can roughly be situated about halfway between the two cluster centers. In this respect the general price variable can indeed be seen as an average price variable that is representative of the price evolution of wheat for the whole territory of France and for the whole period 1756-1870.
- closer analysis of the results reveals that the exact position of the general price variable within the three clusters was changing over time. For the period 1756-1790 the aggregated variable is situated within the second cluster. More precisely, this variable is very close to the center of the second cluster and thus describing the price behavior of wheat in the northern part of France. For the two subsequent periods the aggregated price variable seems to follow very closely the price in the

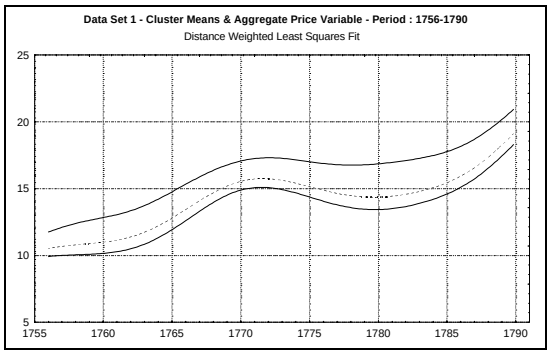
region east. The association between the general price variable and the variable describing the price behavior in the region east must be strong since both variables are switching from the first cluster in period 1797-1830 to the second cluster in period 1831-1870. It follows that for the period 1797-1870 the aggregated price variable for wheat can be seen as describing the price behavior in the region east.

Figure 11 : Wheat Price for Nine Regions - Period : 1756-1870
Original Data Sets - Cluster Means and Aggregated Price Variable
Sequential Line Plot - Distance Weighted Least Squares Fit

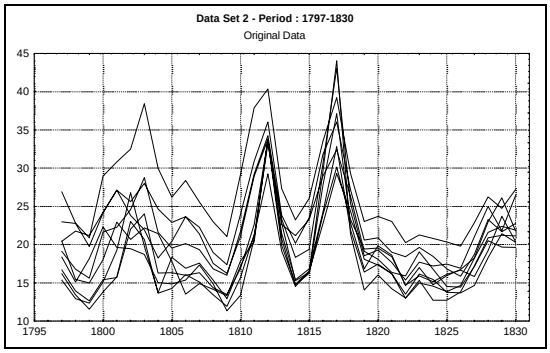
Original Data Sets | **Cluster Means & Aggregated Price Variable**



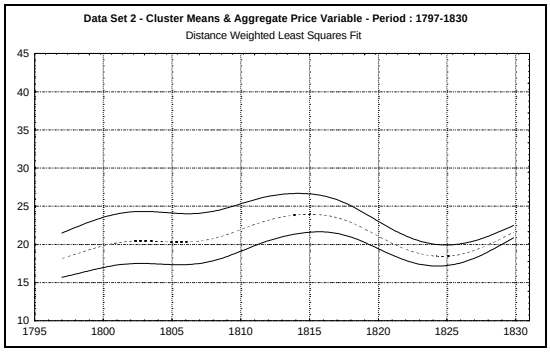
Period : 1756-1790 - Data Set 1



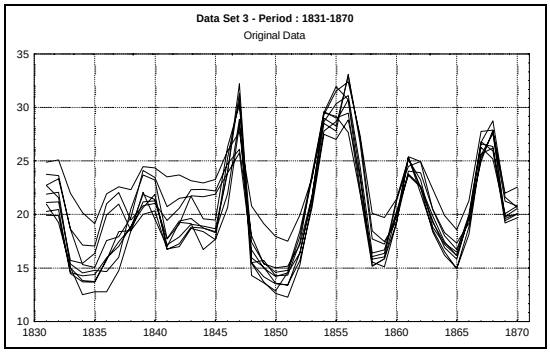
Period : 1756-1790 - Data Set 1



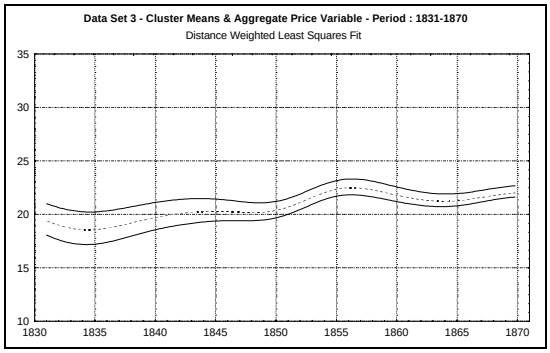
Period : 1797-1830 - Data Set 2



Period : 1797-1830 - Data Set 2



Period : 1831-1870 - Data Set 3



Period : 1831-1870 - Data Set 3

Section 6 : Cluster Analysis - Original Data Transposed

Introduction

So far variables (columns) referred to geographical areas and observations or cases (rows) to the yearly price data for these regions. By transposing the whole data matrix, the geographical areas will become the observations or rows and the years will be the new variables or columns. Consequently, the transposition of the data will result in as many variables as there are years, each of which with as many observations as there are geographical regions.

k-Means Clustering

In this section the k-means clustering method will be applied to the transposed data sets 1, 2 and 3. It must be remembered that the k-means clustering method will produce exactly k different clusters of greatest possible distinction. Each of these k clusters will consist of years for which the distance between the wheat prices of all the regions within the cluster will be smaller than the distance to the other clusters.

Table 7 : Wheat Price for Nine Regions - Period : 1756-1870
Transposed Data Sets - 3-Means Clustering - Cluster Members

Data Set 1				Data Set 2				Data Set 3			
Year	Low	Mid	High	Year	Low	Mid	High	Year	Low	Mid	High
1756	•			1797	•			1831		•	
1757	•			1798	•			1832		•	
1758	•			1799	•			1833	•		
1759	•			1800	•			1834	•		
1760	•			1801		•		1835	•		
1761	•			1802		•		1836	•		
1762	•			1803		•		1837		•	
1763	•			1804	•			1838		•	
1764	•			1805	•			1839		•	
1765	•			1806	•			1840		•	
1766		•		1807	•			1841		•	
1767		•		1808	•			1842		•	
1768		•		1809	•			1843		•	
1769		•		1810	•			1844		•	
1770			•	1811		•		1845		•	
1771			•	1812			•	1846			•
1772		•		1813		•		1847			•
1773		•		1814	•			1848	•		
1774		•		1815	•			1849	•		
1775		•		1816		•		1850	•		
1776		•		1817			•	1851	•		
1777		•		1818		•		1852	•		
1778		•		1819	•			1853		•	
1779		•		1820	•			1854			•
1780		•		1821	•			1855			•
1781		•		1822	•			1856			•
1782		•		1823	•			1857			•
1783		•		1824	•			1858	•		
1784		•		1825	•			1859	•		
1785		•		1826	•			1860		•	
1786		•		1827	•			1861			•
1787		•		1828		•		1862		•	
1788		•		1829		•		1863		•	
1789			•	1830		•		1864	•		
1790			•					1865	•		
								1866		•	
								1867			•
								1868			•
								1869		•	
								1870		•	

By imposing three clusters it can be expected that within each of these clusters the cluster members or years will refer to a specific price level. In other words the final result will consist of a first cluster consisting of the years with low wheat prices for all the geographical areas, a second cluster with those

years characterized by medium wheat prices and a third cluster with those years for which the prices in the regions can be considered as high prices. The exact composition of these three clusters can be found in Table 7.

Results

As could be expected the use of the 3-means clustering technique succeeded, for each of the three data sets, in discriminating between three distinct price levels, i.e. low, medium and high prices. The representation of the composition of the three clusters for each of the three data sets, given by Table 7, can be seen as a highly simplified version of the graphical representation of these data sets given by Figure 12.

Table 8 : Wheat Price for Nine Regions - Period : 1756-1870
Transposed Data Sets - 3-Means Clustering - Cluster Members & Cluster Means

		Cluster Members & Cluster Means					
		Low Prices		Medium Prices		High Prices	
Data	**Period**	**Number**	**Mean**	**Number**	**Mean**	**Number**	**Mean**
Set 1	**1756-1790**	10	11.037	21	15.123	4	19.860
Set 2	**1797-1830**	22	17.824	10	23.926	2	35.112
Set 3	**1831-1870**	13	16.112	18	20.553	9	27.146

In Table 8 even more precise information can be found. Apart from the number of years included in each of the clusters, this table reports also the arithmetic mean of the wheat price of all the members belonging to the cluster . These mean values can be seen as the mean price levels of the clusters or price categories. They were used to construct the graphical representations in the right column of Figure 12. These graphs, representing the price evolution in the nine geographical regions of the three data sets, are overlaid by a step plot. Each of the three levels of this step plot gives the mean level of the relevant price category.

Using the information about the composition of the clusters an even more detailed graphical representation can be constructed. They can be found in the left column of Figure 12. Instead of calculating the overall mean for each of the clusters or categories, a specific and separate mean was calculated for each of the nine regions.

Conclusions

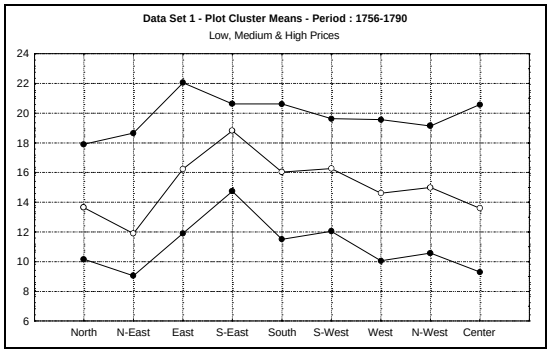
The main characteristics of the previous results can best be summarized by the following conclusions :

- for the period 1756-1790 the medium price level for wheat in the nine regions is situated halfway between the low and high price level
- for the period 1797-1830 and 1831-1870 the difference between the lower and the medium price level is much smaller than the difference between the higher and the medium price level
- for data set 1 the highest prices within both the low and medium price category occurred in the region south-east while the highest prices for wheat in the higher price category can be found in the region east
- the highest price category for the period 1831-1870 is characterized by almost equal mean prices for all nine regions

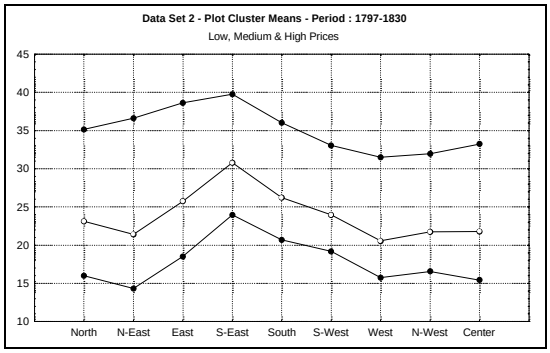
A last remark must be made about this experimental application. In order to discriminate among the observations of time series data, the results obtained by applying the k-means clustering method, illustrates the potential capabilities of this technique as a workable and valuable alternative to the traditional and more specific statistical inferential techniques. The latter do have the disadvantage to be based on even broad hypotheses about the time series generating mechanism, whereas the methodology of clustering is only based on the notion of distance and can even be used for data on the ordinal scale of measurement.

Figure 12 : Wheat Price for Nine Regions - Period : 1756-1870
Transposed Data Sets - 3-Means Clustering - Cluster Means & General Mean

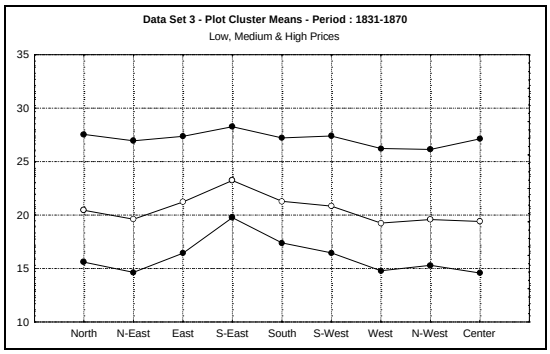
Cluster Means



Period : 1756-1790 - Data Set 1

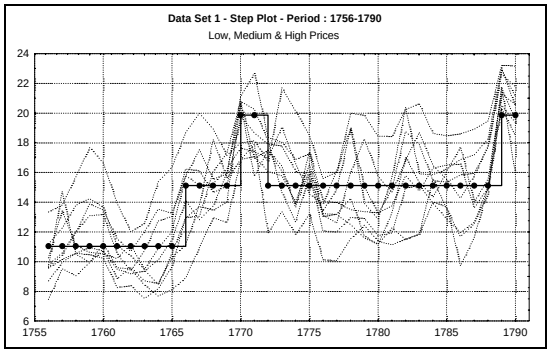


Period : 1797-1830 - Data Set 2

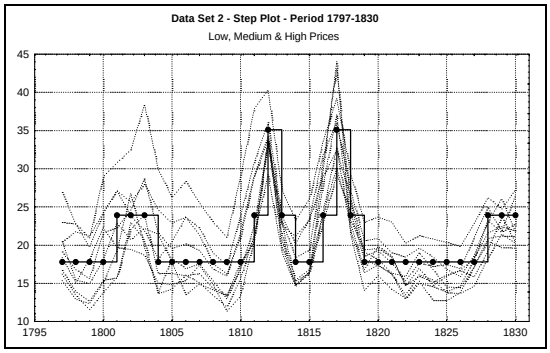


Period : 1831-1870 - Data Set 3

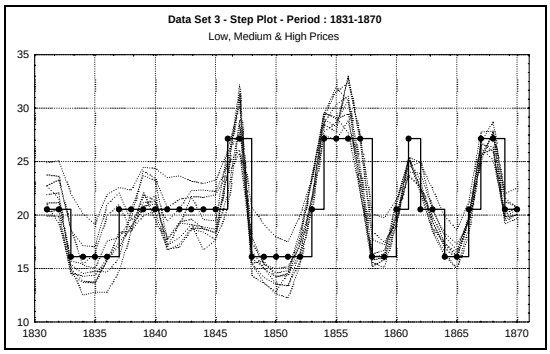
General Mean



Period : 1756-1790 - Data Set 1



Period : 1797-1830 - Data Set 2



Period : 1831-1870 - Data Set 3

Section 7 : Multidimensional Scaling Analysis

Introduction

Multidimensional scaling is another exploratory technique that can be used to find and better understand the information content of a given data set. It can be considered as a valuable alternative to principal component analysis. The main objective of multidimensional scaling is to detect meaningful underlying dimensions that will allow to explain the observed distances, similarities or dissimilarities between variables. The final goal is to find a minimum number of dimensions allowing an almost perfect reproduction of the observed distances, that is, to reduce the observed complexity of nature by explaining the observed distances in terms of a smaller number of underlying dimensions.

There should be no doubt that both Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) can be used to handle about the same research situations. Apart from the similarities between these two multivariate statistical techniques there are, however, fundamental differences between these two methods.

A first difference is that in order to apply the PCA the underlying data must be characterized by linear relationships. MDS does not impose such restriction. The most important among these differences is however that MDS can be used for ordinal data, whereas the level of measurement to apply PCA must be at least at the interval level. A first consequence is that MDS can be used as long as the rank-ordering of distances or similarities of the data is meaningful. In other words, MDS can be applied to any kind of distances or similarities, while PCA requires that a correlation matrix is used to describe the differences or distances between the variables of the data set.

An important and practical consequence of the fundamental difference between these two techniques is that PCA tends to extract more dimensions than MDS. As a result MDS often yields more readily interpretable solutions than PCA. The question remains how many dimensions have to be specified. This is exactly the same problem as the specification of the number of components to be extracted in a principal component analysis. One way to decide how many dimensions to use is to plot an alternative scree test, i.e. the test that was used in the context of the number-of-component problem in principal component analysis. A second criterion for deciding how many dimensions to interpret is the clarity and interpretability of the final configuration.

Several measures can be used to evaluate how well or poorly a particular configuration reproduces the observed similarity or distance matrix. Most of these measures amount to the computation of the sum of squared deviations of the observed distances from the fitted or estimated distances. One of these possible goodness of fit diagnostics is the **Shepard** diagram. In this diagram the actual and observed (normalized) distances are plotted against the reproduced distances, i.e. against the estimated distances according to the monotone regression transformation procedure. The more closely the points in this scatterplot cluster around the diagonal, the better is the fit of the respective model, which then can be considered as adequate in describing the similarities or distances between the nine price series.

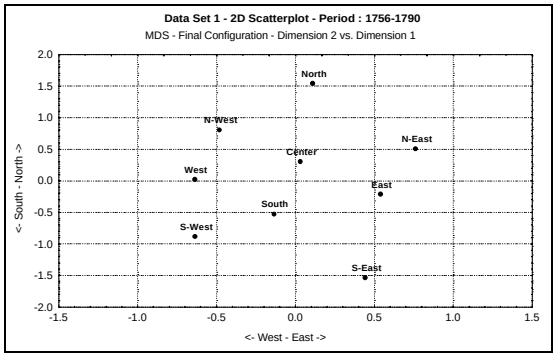
Results

The starting configuration of the algorithm used for this analysis amounts to a principal components analysis of the similarity/dissimilarity matrix. In most instances this will provide an adequate starting configuration for the iterative fitting procedure. Therefore, it was decided to start the MDS analysis by using the correlation matrices for the data sets 1, 2 and 3 and to start the analysis with the computation of a solution for three dimensions. In a second step this solution was used as a starting configuration for a two-dimensional solution. This approach is entirely in line with the MDS philosophy, i.e. to adequately describe the observed distances in terms of the smallest number of dimensions.

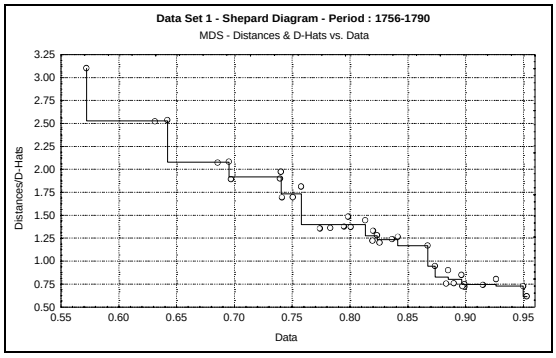
The final results for the two-dimensional solution are given in Figure 13 and Table 9. For each of the three periods 1756-1790, 1797-1830 and 1831-1870, Figure 13 consists of a scatterplot of the final two-dimensional configuration and the **Shepard** diagram. In order to facilitate the comparison of the results for the three periods, the three scatterplots of Figure 13 are represented as one multiple scatterplot in Figure 14. A summary of numerical results is tabulated in Table 9.

Figure 13 : Wheat Price for Nine Regions - Period : 1756-1870
MDS - Final 2D Configuration - Scatterplot and Shepard Diagram

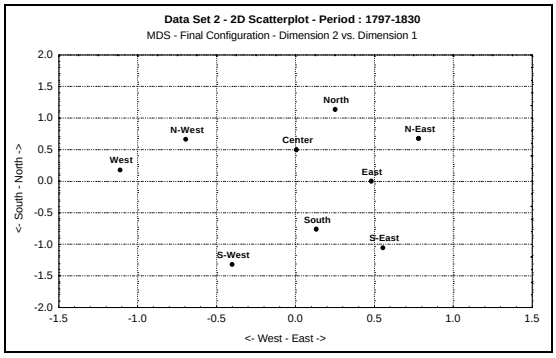
Scatterplot | **Shepard Diagram**



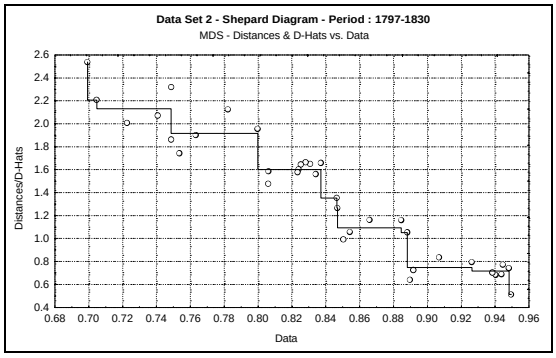
Period : 1756-1790 - Data Set 1



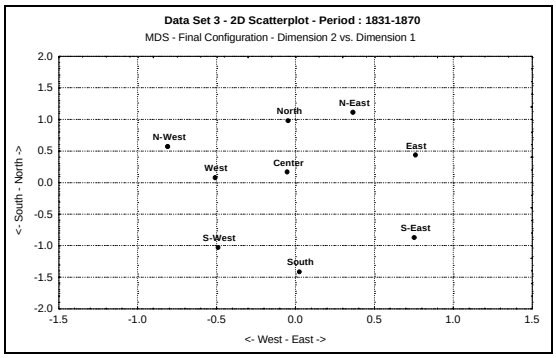
Period : 1756-1790 - Data Set 1



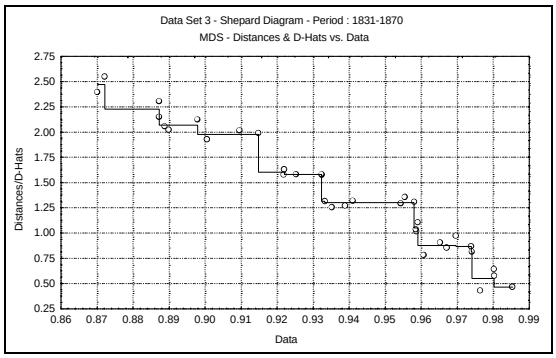
Period : 1797-1830 - Data Set 2



Period : 1797-1830 - Data Set 2



Period : 1831-1870 - Data Set 3

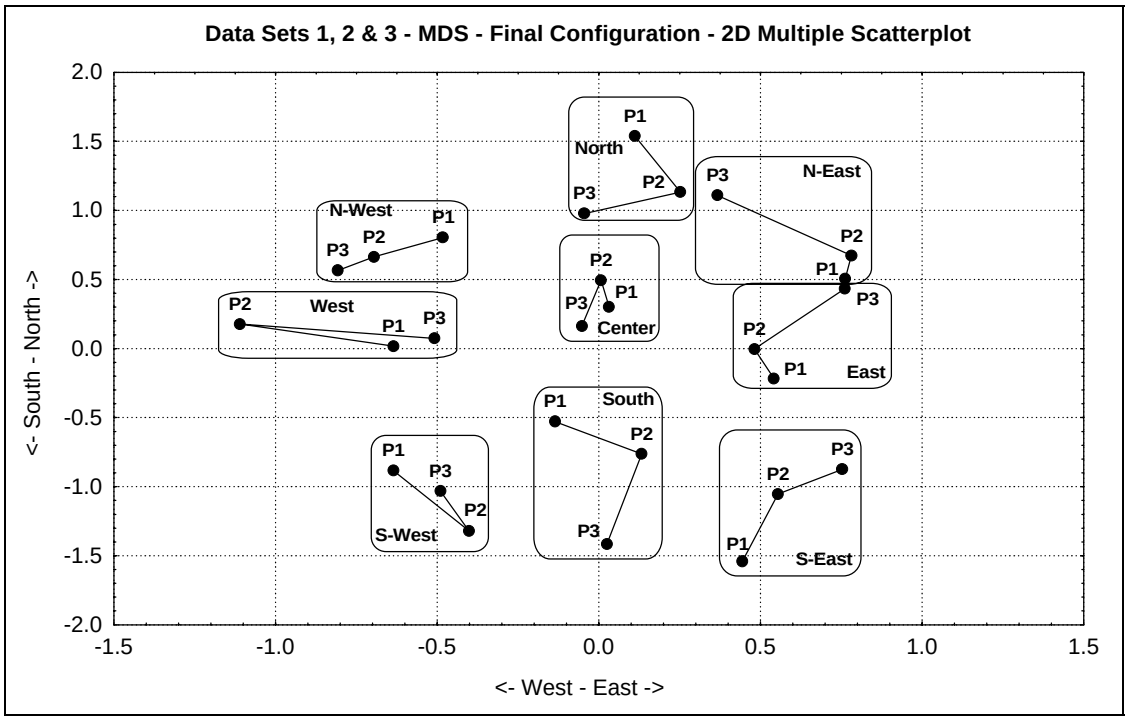


Period : 1831-1870 - Data Set 3

Table 9 : Wheat Price for Nine Regions - Period : 1756-1870
MDS - Summary of Numerical Results - Goodness-of-Fit Tests

	Period		
Statistics	**1756-1790**	**1797-1830**	**1831-1870**
Variables	9	9	9
Observations	35	34	40
Dimensions	3→2	3→2	3→2
Last Iteration Computed	35	30	38
Raw Stress : D-star	0.1457359	0.5269728	0.2768925
Raw Stress : D-hat	0.0532709	0.2108463	0.0890048
Alienation	0.0424075	0.0805931	0.0584423
Stress	0.0256450	0.0510200	0.0331485

Figure 14 : Wheat Price for Nine Regions - Period : 1756-1870
MDS - Final 2D Configuration - Multiple Scatterplot



Conclusions

From all these results the following conclusions can be drawn :

- the decision to retain in the analysis only two dimensions in order to describe the distances or differences between the price variables of the nine regions is largely supported by the diagnostics as there are the goodness of fit tests from Table 9 and the **Shepard** diagrams given in Figure 13
- given the three graphical representations of the final configurations for the two-dimensional solution, presented in Figure 13, these two dimensions can easily be interpreted as the longitude dimension or west-east axis and the latitude dimension or north-south axis
- comparing the final configuration for the three periods, represented by the multiple scatter plot in Figure 14, one can conclude that it is definitely not true that the eight surrounding regions are gradually approaching each other and the center of France
- perhaps the most important conclusion that can be drawn from these results is that the physical geographical distance between the nine regions, expressed in kilometers, seems to be adequate in explaining the difference (similarity or dissimilarity), expressed by the **Pearson** correlation coefficients, between the price variables for these regions

Section 8 : Final Conclusions and Remarks

The main purpose of this paper was to investigate the changing regional differences of the price of wheat in France during the period 1750-1870. By using aggregated price data for nine rather broad defined regions several multivariate statistical descriptive statistical techniques were used. Special attention was paid to present at least the final results of the analysis in a way that would also be accessible by those who are not familiar with the underlying statistical techniques. This explains the abundant use of graphical representations.

From the analyses the following conclusions can be formulated :

- starting around 1810, regional prices are behaving more and more synchronously. Furthermore the period 1810-1870 is characterized by a gradually decline of the difference between the price levels of the nine regions. (Section 1)
- while the largest regional differences of price levels are obtained for the period 1797-1830, these differences are the smallest for the period 1831-1870. (Section 2)
- for the periods 1756-1790 and 1797-1830 the price correlations between the outer regions and the central region as well as the intercorrelations between the outer regions are roughly comparable. For the period 1831-1870 both these correlations are substantially higher. (Section 3)
- for the period 1756-1790 the nine regions can be aggregated into three distinct clusters or groups, i.e. a first group consisting of the three southern regions, a second group with the western and eastern region and a third group clustering the remaining four regions. For the period 1831-1870 the regions of the second group are joining the third cluster, resulting in a two-cluster situation. In this respect the period 1797-1830 can be seen as a transition period. (Section 4)
- by using appropriate clustering techniques, the difference between the price behavior in the three southern regions and those in the remaining regions is confirmed for each of the three subperiods. (Section 5)
- the potential capabilities of using clustering techniques for the analysis of the transposed data are illustrated for each of the three sub-periods. (Section 6)
- the most important conclusion that can be drawn from using the MDS-technique is that the difference (similarity or dissimilarity) between the wheat prices in the nine regions, expressed by the **Pearson** correlation coefficients, seems to be adequate in explaining the physical geographical distance between these nine regions. (Section 7)

The main theme of this paper has been the presentation of the results from the analyses of the price behavior of wheat in nine regions of the French territory. Since only nine regions were used it is evident that only general conclusions could be drawn from such an analysis. In a next paper the same periods will be analyzed by using more disaggregated price data. It is hoped that this would lead to the refinement of the promising results obtained in this paper.

Appendix 1 : Data

1756<-	Set 0	->1870

1756<-	Set 1	->1790	1797<-	Set 2	->1830	1831<-	Set 3	->1870

Data Set 0 : Prix Moyens Nationaux Annuels de Froment - Par Année Civile

Source : See **Labrousse et al. [11, pp. 9-11]**
Period : 1756-1870 (1726-1913)
Missing data for the period 1793-1796
Regions : France
Series : 1
Frequency : Yearly
Observations : 111 (184) per series
Price / Unit : Livres et centièmes de livre / Hectolitre

Data Set 1 : Prix Moyens Interrégionaux Annuels de Froment - Grands Secteurs Territoriaux

Source : See **Labrousse et al. [11, p. 23]**
Period : 1756-1790
Regions : Grands Secteurs Territoriaux - See Appendix 2
Series : 9
Frequency : Yearly
Observations : 35 per series
Price / Unit : Livres et centièmes de livre / Hectolitre

Data Set 2 : Prix Moyens Interrégionaux Annuels de Froment - Grands Secteurs Territoriaux

Source : See **Labrousse et al. [11, pp. 23-24]**
Period : 1797-1830
Regions : Grands Secteurs Territoriaux - See Appendix 2
Series : 9
Frequency : Yearly
Observations : 34 per series
Price / Unit : Livres et centièmes de livre / Hectolitre

Data Set 3 : Prix Moyens Interrégionaux Annuels de Froment - Grands Secteurs Territoriaux

Source : See **Labrousse et al. [11, pp. 27-35]**
Period : 1831-1870
Regions : Grands Secteurs Territoriaux - See Appendix 2
Series : 9
Frequency : Yearly
Observations : 40 per series
Price / Unit : Livres et centièmes de livre / Hectolitre

Appendix 2 : Grands Secteurs Territoriaux

#	Region
1.	North
2.	North-East
3.	East
4.	South-East
5.	South
6.	South-West
7.	West
8.	North West
9.	Center

Remark

In **Labrousse et al. [11 , p. 21]** the following information about these regions is given :

'…Grands Secteurs Teritoriaux…Secteurs constitués au XIXe siècle par les services nationaux de la statistique agricole. On s'est efforcé de grouper les généralités de l'ancien régime dans le cadre de ces secteurs…'.

References

General

[1] Cox, T.F. & Cox, M.A.A.
Multidimensional Scaling
Chapman & Hall, London, 1994.

[2] Crook, M.
Revolutionary France : 1788-1880
The Short Oxford History of France
Oxford University Press, Oxford, 2001.

[3] Doyle, W.
Old Regime France : 1648-1788
The Short Oxford History of France
Oxford University Press, Oxford, 2001.

[4] Everitt, B.S.
Cluster Analysis
Edward Arnold, London, 3rd Ed., 1995.

[5] Everitt, B.S. , Landau, S. & Leese, M.
Cluster Analysis
Edward Arnold, London, 4th Ed., 2001.

[6] Hakstian, A.R. , Rogers, W.D. & Cattell, R.B.
The Behavior of Numbers of Factor Rules with Simulated Data
Multivariate Behavioral Research, vol. 17, 1982, pp. 193-219.

[7] Jackson, J.E.
A User's Guide to Principal Components
John Wiley & Sons Inc., New York, 1991.

[8] Jolliffe, I.T.
Principal Component Analysis
Springer, New York, 2nd Ed., 2002.

[9] Ward, J.H.
Hierarchical Grouping to Optimize an Objective Function
Journal of the American Statistical Association, vol. 58, March, 1963, pp. 236-244.

[10] Zwick, W.R. & Velicer, W.F.
Comparison of Five Rules for Determining the Number of Components to Retain
Psychological Bulletin, vol. 99, 1986, pp. 432-442.

Price Data

[11] Labrousse, C.-E., Romano, R. & Dreyfus, F.-G.
Le Prix du Froment en France
Au temps de la Monaie Stable, 1726-1913
Ecole Pratique des Hautes Etudes - VIe Section
Centre de Recherches Historiques
Monnaie - Prix - Conjoncture IX
SEVPEN, Paris, 1970, pp. 9-11, pp. 23-24 & pp. 27-35.

Software

[12] Statistica
Kernel Release 5.5, Edition ’99
Statsoft, Tulsa, 1999.