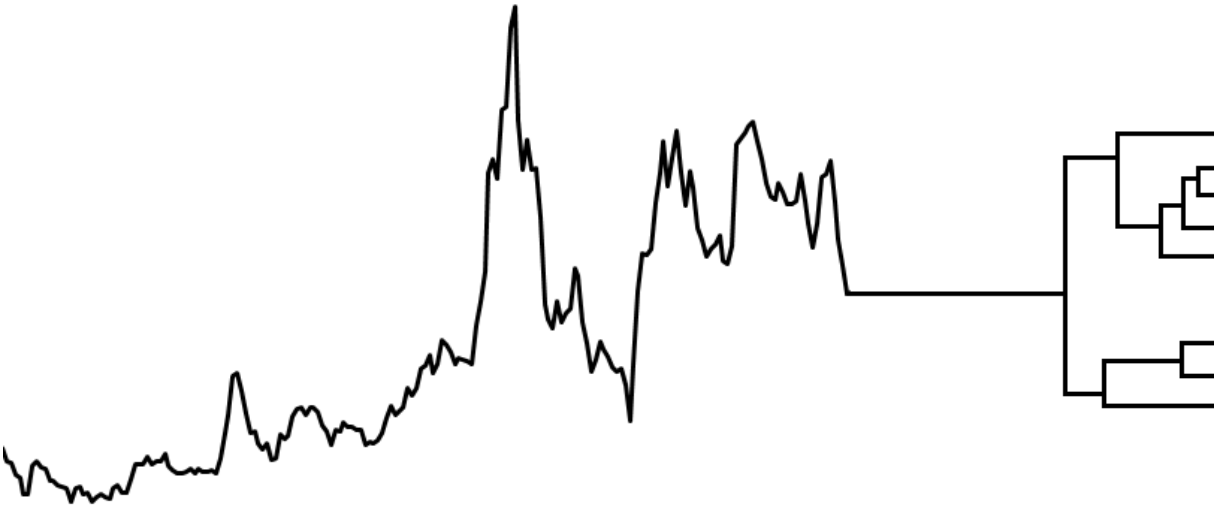
Analysing market integration

An alternative approach



MSc Thesis (AEP-80433)

Student
M. Kind (MID - 871125435050)
Supervisors
Dr. ir. C. Gardebroek
Dr. R. Ihle

17 March 2015
Wageningen

WAGENINGEN UR
For quality of life

Acknowledgement

This thesis was written for the completion of the Master International Development Studies at Wageningen University. I worked on it from November 2014 to March 2015 at the Agricultural Economics and Rural Policy (AEP) group. I would like to extent my gratitude to my supervisors, Dr. ir. C. Gardebroek and Dr. R. Ihle, for introducing me to the central idea of this thesis, providing valuable feedback and the highly enjoyable discussions throughout the process. Lastly, I would like to thank my parents for always having my back and believing in me.

Table of Contents

Abstract

Market integration is an important concept for agricultural economics. The extent to which a local market is integrated in the wider regional or even global market is indicative of possible arbitrage opportunities and whether price changes from international markets are transmitted to the local market. A popular econometric technique that tests the degree of market integration is cointegration analysis. However, when considering data that has serious shortcomings, the exact parameter estimates obtained through cointegration quickly lose their value. The technique may even give the researcher a false sense of exactness, which is not supported by the data. For data from developed countries this shortcoming is, in general, not particularly restricting. Data from developing countries, on the other hand, is more often found to be lacking and missing observations, changing data frequencies, outliers and measurement error are far more common. The purpose of this thesis is to investigate whether cluster analysis is a viable alternative to the cointegration approach for investigating market integration. Both methods are applied to a set of generated time series with known properties. Results indicate that cointegration analysis struggles both with deviations from the ideal market relationship and analysing markets in a multivariate framework. Cluster analysis is shown to provide accurate results when analysing a large number of markets simultaneously. Additionally, cluster analysis is subjected to a sensitivity analysis and is shown to be quite robust to missing observations in the data, with the final clustering outcome only being affected to a marginal degree with 20% missing observations. Cluster analysis is not able to provide information on short and long term effects or adjustment speed. When analysing markets in a bivariate setting with high quality data, cointegration is thus the preferred method. However, when analysing large numbers of time series simultaneously that potentially feature seriously flawed data, cluster analysis provides a valuable alternative to cointegration analysis in empirical market integration research. Further research is needed to determine an optimal set of basic variables to be used for cluster analysis and reduce the subjectivity of the method by developing a standard stopping rule.

1. Introduction

1.1 Problem definition

Market integration is an important concept in international economics. The extent to which a local market is integrated in the wider regional or even global market is indicative of possible arbitrage opportunities and whether price changes from international markets are transmitted to the local market. Should the local market be fully integrated in the global market, then we can expect the so called law of one price (LOP) to hold in the long term. Arbitrage between low-price exporting regions and high-price importing regions ensures the price difference between regions is reduced to the marginal transportation and handling costs (Vercammen 2011). Fackler and Goodwin (2001) point out that supportive evidence for the LOP exists, particularly when considering the long-term, tradable goods and when accounting for transaction costs. Market integration is important for farmers in developing countries that are trying to connect to larger international markets. The degree of price transmission is also vital in understanding the impact of global food price spikes on local markets. If there is a negligible degree of integration, local food consumers will hardly be affected by price shocks in other markets. Producers are also affected by the degree of market integration, since they base their production decisions on expected and observed price levels.

Rapsomanikis et al. (2006) suggest that the absence of market integration has important consequences for economic welfare. A low degree of price transmission, caused by trade policy or due to transaction costs caused by a lacking physical and communication infrastructure, reduces the price information available to economic agents. The immediate result is incomplete information and this could potentially lead to inefficient market outcomes. Contentious issues, such as trade liberalisation and the distribution of costs and benefits across societies, are thus fundamentally affected by the degree of market integration. Hence it is vital that this concept is properly understood in order to develop effective policy.

Governments can severely disturb the degree of market integration by implementing border policies or price support mechanisms. An import tariff for example, if it is set prohibitively high, can potentially result in anything from partial price transmission to a complete breakdown of price transmission. The domestic and international prices can start moving independently from one another, as in the case of an import ban. High transfer costs can also have an insulating effect on the domestic economy. This is particularly important in developing countries, where poor infrastructure can strongly increase the costs of delivery for both imported and exported goods. This reduces arbitrage opportunities and has a direct impact on the degree of price transmission. Another potentially hindering factor is non-competitive behaviour by firms. Oligopolistic tendencies and collusion among firms may result in price differences that are above the marginal transportation and handling costs. (ibid)

A popular econometric technique that tests the degree of market integration is cointegration. It can be used in a wide range of different markets. Recent examples for the wheat market include an assessment of price dynamics between Ukraine, Russia, the United States, the European Union and Canada (Goychuk & Meyers 2014), a study on the impact of the Uruguay Round Agreement on Agriculture on wheat prices in Pakistan (Mukhtar & Ishfaq 2009), an analysis of the price transmission

between Germany and Hungary (Bakucs et al. 2012) and an in depth look at integration between wheat markets in India (Ghosh 2003).

Granger (2004), the man who first coined the term cointegration, explains the term in a straightforward manner by stating that if it turns out that the difference between a pair of integrated series is stationary, these series are said to be cointegrated. When price series of different markets are cointegrated this signals a stable long-run relation between these series, even if both series wander due to deterministic or stochastic trends. Essentially the main point to remember is that when series are cointegrated, they share a common trend that ensures prices develop in a similar fashion over time (Verbeek 2012). Price series that are cointegrated can be modelled together in a vector error correction model. This particular type of model describes how series behave in the short-run, while being consistent with a long-run cointegrating relationship (ibid).

The main drawbacks of the cointegration approach are that extensive time series data has to be available and that this data has to be consistently measured in order to avoid making incorrect inferences. This implies that the time series have to be sufficiently long and complete (i.e. no missing observations), since the approach requires the use of lags and differences of the time series variables. For data from developed countries these criteria are not particularly restricting. This is not to suggest that data related problems are absent in time series from developed countries, but to emphasize that problems are less frequent and less severe than in other parts of the world. Data from developing countries, on the other hand, fail to meet these criteria more often and missing observations, changing data frequencies, outliers and measurement errors are far more common. Furthermore, in these countries the problem of missing observations may be tackled by simply interpolating last period's value. Lastly, time series from developing countries can be too short or feature too few observations per sub-period. To properly investigate market integration in these countries one may have to resort to other quantitative methods, which are more robust against ill-posed data.

1.2 Research objective

The purpose of this thesis is to investigate whether cluster analysis (see Hair et al. 1998 or Lattin et al. 2003) is a viable alternative to the cointegration approach for investigating market integration. Cluster analysis focusses on categorization and does this by categorizing a large number of groups into a smaller number of sub-groups. Observations within the same group will be relatively similar, while observations placed in distinct groups will be relatively dissimilar. This potentially makes cluster analysis an ideal research tool when investigating the degree of market integration between large numbers of markets simultaneously. The exact measure of similarity is for the researcher to decide; several different alternatives for the metric and the linkage criterion are available and will be discussed. Market integration will be investigated by defining a number of statistics describing the evolution of a price series (e.g. mean, variance, skewness and kurtosis) and applying a cluster analysis on these non-parametric statistics. The crux is whether or not this analysis will yield clusters of markets that were also found to be integrated when using the cointegration approach. Furthermore, to investigate the robustness of cluster analysis a sensitivity analysis will be conducted. This will reveal how strongly the results of cluster analysis are affected by missing data.

1.3 Research questions

In order to achieve the research objective specified above and ascertain whether cluster analysis is indeed a viable alternative to cointegration for market integration research, the following four research questions will be answered within this thesis.

1. What is market integration and what are its defining properties?
2. How does the cointegration approach function and which markets are found to be integrated in the data at hand?
3. How does cluster analysis operate and which markets are found to be integrated in the data at hand?
4. Do cointegration and cluster analysis lead to similar conclusions and how robust is cluster analysis to data problems?

1.4 Methodology

In order to answer the first research question a literature review will be conducted. The second and third research questions will be answered by applying the cointegration approach and cluster analysis to the data used in this study, respectively. The main time series data has been obtained from the World Bank's GEM Commodities database (http://data.worldbank.org/data-catalog/commodity-price-data). The remaining eleven time series used in the analyses are artificially generated, using the world price as a basis. The utilized data is provided in nominal US dollars per ton and the commodity of interest is wheat, since this crop is one of the worlds most traded agricultural commodities (FAO 2014).

1.5 Outline

What follows is the presentation of the answers found to the research questions. The next chapter serves to introduce the theoretical framework, which is central in developing the theoretical understanding of what market integration entails. The third chapter focusses on the methodology and provides the theoretical underpinnings of cointegration and cluster analysis. Moreover, it also features a brief discussion that compares cointegration with cluster analysis. Chapter four introduces the data and clarifies how the different time series were generated. In the fifth chapter the results of the analyses are reported for both cointegration and cluster analysis. Furthermore, the sensitivity analysis is featured within this chapter. The final chapter makes a direct comparison between the results of both aforementioned data analysis methods and features the primary conclusions of the thesis. Since the application of cluster analysis as a market integration research tool is a novel approach, some suggestions for further research will also be supplied here.

2. Theoretical framework

For developing countries the agricultural sector is a crucial one, as a large part of the total workforce is often employed in the sector and it can contribute to development through exports (Perkins et al. 2006). By integrating into a larger international agricultural market developing countries can expand their export opportunities, while internal supply shocks can also be dealt with more appropriately. To properly analyse market integration it is important to first gain a meaningful understanding of the concept itself. This chapter focusses on building this understanding by reviewing the properties of market integration, discussing the major limiting factors that inhibit market integration, highlighting the development of empirical methods used to analyse market integration and briefly considering four recent articles that focus specifically on market integration for the commodity wheat.

2.1 Properties

The degree to which domestic agricultural commodity markets respond to price changes on the global market is central in understanding the degree of market integration of an economy (Rapsomanikis et al. 2006). A fundamental part of this is the degree of price transmission, which is the degree of co-movement that can be observed of a commodity price for a region in response to changes in another region. Or in other words, the focus of this thesis is on spatial price integration of spot prices in different regions. Horizontal and vertical price integration are not considered. Horizontal price integration is the extent to which a price change in one commodity affects the price of another commodity. Vertical price integration focusses on the supply chain and reveals the effect of a price change on producers, processors, retailers and consumers. (Vercammen 2011)

Using the terminology of the Law of One Price, one can argue that there is complete price transmission if the equilibrium prices of a commodity in two different markets differ only by the marginal transportation and handling costs between the regions. Spatial arbitrage through trade will ensure that changes in supply and demand in any of the two markets will not violate this law in the long run. This is illustrated by equation 2.1 (ibid), which simply states that prices in region 1 (P_1) are equal to prices in region 2 (P_2) plus some form of transaction costs ($T_{2,1}$). The concept of transaction costs is broader than merely transportation and handling costs, also including tariffs, information costs, insurance premiums, negotiation costs, etc. (Barrett 2001; Barrett & Li 2002; Abdel-Latif & Nugent 1996).

$$P_1 = P_2 + T_{2,1} \qquad (2.1)$$

In this particular example region 2 is exporting a commodity to region 1. The equality suggested by this equation is known as the strong form of the Law of One Price (Fackler & Goodwin 2001). However, in reality this will not always hold. Should the price in region 1 be lower than the price in region 2 plus transaction costs ($P_1 < P_2 + T_{2,1}$), no trade will occur since the gains from trade will be negative. On the other hand, if the price in region 1 is higher ($P_1 > P_2 + T_{2,1}$), the gains from trade are positive and arbitrageurs will be able to profit from the situation. They will increase shipments from region 2 to region 1, increasing the price in region 2 (due to a reduction in supply) and decreasing the price in region 1 (due to an increase in supply). Theoretically this process will continue until the gains are exhausted and the equilibrium condition specified in equation 2.1 is restored.

To measure the degree of market integration Fackler & Tastan (2008) and Fackler & Goodwin (2001) focus on the price transmission ratio, which measures to what extent a price shock in one region is transmitted to another region. Suppose there is a shock (ε_1) that shifts excess demand in region 1, while region 2 is unaffected. The price transmission ratio ($R_{1,2}$) corresponding with this shock can be calculated by using equation 2.2.

$$R_{1,2} = \frac{\partial P_2 / \partial \varepsilon_1}{\partial P_1 / \partial \varepsilon_1} \tag{2.2}$$

This ratio is not necessarily symmetric (i.e. $R_{1,2} \neq R_{2,1}$), implying that region 1 can be more integrated with region 2 than region 2 is with region 1. For the example illustrated above the price transmission ratio is either 0 or 1. This is due to the fact that in this simple model the transaction costs per unit are not sensitive to the amount shipped, hence they are constant. Transaction costs are said to be perfectly elastic; the amount shipped has no influence on per unit cost. Under this assumption one can argue that if there is no trade, markets are not connected and small changes in price are not transmitted between the two regions. Hence, in this situation prices in both markets will fluctuate independently and will not be related. If there is trade, however, markets are connected and changes in price are fully transmitted between regions. Barrett (2001) notes that transaction costs between two markets need not be stationary, due to non-linear transaction costs, while still maintaining spatial equilibrium between the two regions. This obfuscates the analysis of market integration.

The example above is a rather strong oversimplification of reality and more complex spatial models do exist. As is to be expected, the price transmission ratio does not only take values of 0 and 1 in reality, but can take any value between these two extremes. Fackler & Tastan (2008) use the so called Takayama-Judge model, which incorporates a measure for transaction cost supply elasticity to allow for the price transmission ratio to vary between 0 and 1. They subsequently use this model and the price transmission ratio to derive three distinct ways of measuring market integration. The first two measure the degree of integration between pairs of regions, while the third measures the degree of integration in the entire market. It is noted by these authors that market integration can be due to both direct and indirect trade. If two regions do not directly trade with each other, they can still be integrated if they are part of a common trading network. Hence, it is important to keep the complete market in mind when describing the degree of market integration and not focus exclusively on pairwise combinations of regions.

Prices between regions can be related in a plethora of ways and in order to reduce this complexity Rapsomanikis et al. (2006) base their conception of price transmission between regions on the following three components:

- Co-movement and complete adjustment of prices, which implies that price changes in one market are completely and instantaneously transmitted to the other market.
- Price dynamics and speed of adjustment, which indicates the process and rate of price adjustment from one market to the other in the short and long run.
- Asymmetric price response, implying that both the extent and the speed of price adjustment between markets can be transmitted symmetrically or asymmetrically.

These components further illustrate that market integration is not limited to the situation of complete price transmission, where changes are completely and instantaneously transmitted between markets. Price transmission can be incomplete in the short run, while still being complete in the long run. In this case markets are integrated, but there is a temporal effect or delay and the speed of adjustment plays a role. Asymmetric response to a price change also does not imply that markets are not integrated. Rapsomanikis et al. (2006) point out that in the literature market power is often seen as the primary cause of asymmetries. For example, a monopsonist might be inclined to slowly transfer price increases in the international market to its suppliers, while instantaneously transferring price decreases. This implies that the degree of market integration is directly affected by asymmetric price responses. Meyer & Cramon-Taubadel (2004) develop asymmetry more fully by conceptualizing it based on three criteria. The first criterion highlights the magnitude and speed of price transmission. The magnitude of the price response can differ depending on whether it is a price increase or a price decrease. Following the same logic, the speed of adjustment can also differ depending on whether prices increase or decrease. The second criterion classifies asymmetric price transmission as either positive or negative. If the price of wheat in region 1 reacts more fully, in terms of magnitude and speed, to a price increase than a price decrease in region 2, the asymmetry is said to be positive. The reverse holds in the case of negative asymmetry. The third criterion refers to whether the asymmetric price transfer is vertical or spatial. In addition to market power, Meyer & Cramon-Taubadel (2004) add adjustment costs as a major factor contributing to asymmetric price transmission. For spatial asymmetric price transmission this can arise due to varying transportation costs depending on the direction of trade (e.g. trade flows up and down a river).

When studying spatial price changes, it is also important to keep in mind the distinction between competitive equilibrium and market integration (Barrett & Li 2002). Spatial market relationships are commonly described by their prices, as is done in the remainder of this thesis, by trade volumes or a combination of the two. Both of these indicators have their advantages and disadvantages. Price analysis does not reveal anything about the actual trading behaviour of regions, while trade volumes cannot reveal whether competitive equilibrium conditions hold. Barrett & Li (2002) attempt to combine these two measures into one technique that is able to distinguish between competitive equilibrium and market integration. The former is defined as the situation where competitive pressures drive the marginal rents from trade down to zero, while the latter reflects the tradability of goods between spatially distinct regions.

Based on the above definition these two distinct concepts are thus characterized by rents and trade volumes. Rents can be positive, negative or zero, while trade volumes can either be zero or positive. Hence these two characterizing variables allow for 6 distinct situations, which the aforementioned authors divide into the four categories specified in table 2.1.

Table 2.1: Categorizing equilibrium and integration (Barrett & Li 2002)

Market condition	Rent	Trade volume	Related concept
Perfect integration	0	≥ 0	Competitive equilibrium & market integration
Segmented equilibrium	< 0	0	Competitive equilibrium
Imperfect integration	$\neq 0$	> 0	Market integration
Segmented disequilibrium	> 0	0	-

Perfect integration implies that rents are reduced to zero, while trade volumes can either be zero or positive. When rents are equal to zero, two market are still said to be perfectly integrated even when trade volumes are zero, because in this instance arbitrageurs are simply indifferent to trading. Perfect integration combines the concepts of competitive equilibrium and market integration, since rents are reduced to zero and tradability of goods is present. Segmented equilibrium holds when potential rents are negative and there is no trade being conducted. This situation only reflects the concept of competitive equilibrium. Under conditions of imperfect integration, rents can be either positive or negative, while trade volumes will always be positive. In the case of positive rents either arbitrage opportunities are being left unexploited by traders or a significant part of the transaction costs are unobservable. When profits are negative, on the other hand, there could be some form of temporary disruption or significant unobservable transaction benefits. Both instances of imperfect integration are consistent with the concept of market integration. Lastly, in the case of segmented disequilibrium there is a potential for positive rents, but trade volumes are still zero. This could be due to, for example, a trade ban. This market condition does not comply with either of the two concepts.

The above categorization is important due to the fact that it disentangles the concepts of competitive equilibrium and market integration, which allows for a more accurate analysis of actual market integration. Barrett & Li (2002) point out that existing methods classify segmented equilibrium as integrated, despite the lack of price transmission and the absence of trade flows. Furthermore, imperfect integration is often classified as not being integrated, despite the fact that trade is occurring under this market condition.

2.2 Limiting factors

As was discussed in detail in the previous paragraph, price transmission between markets is not necessarily complete under real world conditions. On the contrary, complete transmission is somewhat of a rarity in the real world, especially in developing countries. Minot (2011) highlights six factors that reduce or slow price transmission from one market to another. These six factors will be discussed in order, supplemented with additional information from other sources. It should be noted that the measurement of price integration receives considerable attention in the literature, while the determinants of integration are rarely discussed (Fackler & Goodwin 2001; Varela et al. 2012).

Firstly, imported and local goods are often considered to be perfect substitutes, as if they are a homogenous product. In other words, maize is considered to be maize regardless of its origin. If this is true, vendors cannot charge a different price based solely on the origin of the commodity and the price of the local and imported good is exactly the same. However, consumers often have a preference for one commodity over the other based on some perceived difference in quality. The products can thus be imperfect substitutes and prices for both commodities may differ, implying an imperfect degree of price transmission. A related point of interest is that consumer preferences are not homogenous and a wide range of staple foods could potentially be part of a local diet. Suppose that international prices of wheat experience a significant upward shock. Consumers in a region that imports part of its wheat supply may simply switch to substitutes, such as rice or maize. This will limit the upward price transmission of wheat, while consequently resulting in an increase in price of the other local staples. The effect of the price shock is thus spread across different local markets (Cudjoe et al. 2010).

Secondly, if a small number of traders dominate a market they may have a substantial degree of market power, allowing these traders to influence prices to a degree. They could, for example, be inclined to quickly transmit price increases, while being reluctant to transmit reductions in price. Rapsomanikis et al. (2006) also find this limiting factor in the literature and suggest that oligopolistic behaviour and collusion between firms can strongly disrupt price transmission. This particular kind of non-competitive behaviour thus results in a gap between international and domestic prices that cannot solely be contributed to transaction costs.

Another limiting factor is the lack of perfect information that is often assumed in spatial economic models. If traders and producers are not instantly aware of price changes in world markets, they are less able to respond to potentially profitable opportunities. Producers will have a weaker negotiation position, while traders are not fully able to execute their roles as spatial arbitrageurs. This may result in economic agents taking decisions that contribute to inefficient economic outcomes, which has negative implications for economic welfare (Rapsomanikis et al. 2006).

Furthermore, trading does not occur instantaneously and can take a substantial amount of time. Once a trader decides to import a commodity from overseas, the shipment has to be prepared, loaded and offloaded several times, transported by boat and possibly train or truck, go through customs etc. Due to this process, price transmission is slowed and differences in price can persist over time before being corrected by spatial arbitrage. Long term commitments that are formalized in the form of a contract can also delay transmission of price signals (Barrett & Li 2002). A recent example is the Russian boycott of EU foods, resulting in a significant price drop for producers. However, consumer prices in supermarkets did not immediately follow suit, since retailers had long-term contracts with their suppliers (Volkskrant 2014).

The fifth limiting factor is trade policy. In the standard view of most international economists free trade is the ideal that should be strived for and governments should interfere with trade as little as possible. The standard argument in favour of free trade stresses efficiency improvements and a reduction of rent seeking behaviour that is often the result of governments implementing trade policy (Krugman et al. 2012). However, in reality governments can and do implement a myriad of trade policies, such as tariffs, import subsidies and quotas. Import tariffs increase transaction costs, but do not directly affect price transmission. When set prohibitively high, however, opportunities for spatial arbitrage are eliminated and the world and domestic price levels will move independently as if an import ban has been implemented. Moreover, quantitative restrictions can cause price transmission to completely break down. Other policies, such as trade licenses, tariff rate quotas or price floors, all have similar effects in that they result in incomplete price transmission and reduce adjustment speed. Furthermore, if governments intervene in trade in a sporadic, seemingly random, manner, this increases the risk to traders and can discourage them from participating in international markets. Hence, trade policy has numerous and complex consequences for the degree of price transmission and market integration of an economy (Minot 2011; Rapsomanikis et al. 2006).

Transaction costs, particularly transportation costs, are the final limiting factor and are a major element in trade patterns. Commodities in particular have a low value-to-bulk ratio, which implies that transfer costs are relatively high when compared to the price of the commodity itself. In less developed regions of the world, often with poor physical infrastructure, the cost of delivering the commodity to local markets can be very substantial. Minot (2011) claims that for imported grains in

Sub-Saharan Africa, these costs can constitute up to half of the final price paid by consumers. This may ultimately hinder price transmission, since arbitrage is strongly inhibited. Consequently, world market price changes are not fully, if at all, transmitted to local domestic markets and economic agents will only partially adjust to shifts in world supply and demand (Rapsomanikis et al. 2006).

2.3 Empirical Methods

Market integration is a concept that has multiple dimensions and is thus difficult to measure empirically. It is likely impossible to give a universal measure of the extent of price transmission using only a single parameter or statistical test (ibid). The economic literature has focussed primarily on econometric time series analysis of prices. This method relies exclusively on price data, which are typically more readily available for developing countries. Time series are able to reveal co-movement of prices, but also allow the researcher to distinguish asymmetries and magnitudes. Over the years several techniques have been applied to analyse these time series, becoming increasingly sophisticated and more powerful. Economists have moved from bivariate correlation coefficients, to static regression analysis and finally to the cointegration approach. These three methods will be briefly discussed in this section. Chapter three features a more comprehensive explanation of the cointegration approach, which has become the standard tool for analysing spatial market relations.

Bivariate correlation coefficients were commonly used in the early days of market integration research. With this method higher correlation coefficients, that were statistically significant, would be indicative of market integration. The method, however, is not able to capture dynamic relations among markets and is sensitive to so called spurious relationships. Spurious correlations can occur when a third factor, such as inflation, influences both variables under consideration in a similar manner. This results in high correlations, suggesting a degree of market integration, while there is no direct causal relationship between the two variables (Sharma 2003).

Static regression analysis was the next econometric method to be applied to study market integration. Equation 2.3 shows the simplest specification for this method.

$$P_t^d = \beta_1 + \beta_2 P_t^W + \varepsilon_t \qquad (2.3)$$

In this equation P_t^d is the domestic price in period t, P_t^W is the world price, β_1 and β_2 are model parameters to be estimated and ε_t represents the error term and is assumed to be $IID(0, \sigma^2)$. When applying this method to price data that is expressed in natural logarithms, β_2 is a direct measure for the price transmission elasticity. However this technique suffers from serious problems, including non-stationarity of the data and high degrees of positive autocorrelation leading to spurious regression results, and is also no longer used for market integration research. (ibid)

In order to overcome the weaknesses of the methods mentioned above, modern time series econometrics is based on dynamic regression models. Chief among these methods is the cointegration approach used in error correction models. Other methods, such as dynamic regression based on point location models, Granger causality and the Ravallion market integration criteria, will not be considered here. Cointegration tests the long run relationships in dynamic systems, which is vital considering the dynamic nature of commodity trade in terms of temporal lags. In essence, cointegration between time series is said to occur when the series themselves are non-stationary,

but here exists a linear combination of the series that is stationary. The individual series may wander apart, seemingly without any significant relationship between them, for short periods of time, but they will not wander too far apart in the long term. It is possible to both test direct bivariate cointegration between pairs of markets and to evaluate larger markets, consisting of n series, where the number of cointegration relationships found is considered an indication of overall market integration (Fackler & Goodwin 2001).

Cointegration thus tests long run tendencies, rather than period by period price responses. Theory suggests prices will not drift too far apart due to spatial arbitrage; hence the price spread should be stationary. The implicit assumption here is that transaction costs are in fact stationary, while there is a plethora of eventualities that can seriously challenge and even completely invalidate this assumption (e.g. fluctuations in the price for oil) (Fackler & Goodwin 2001; Rapsomanikis et al. 2006). When price spreads are stationary, perfect integration between markets is straightforward to detect with the cointegration approach. However, as discussed in section 2.1, there are also situations in which markets are imperfectly integrated and correct interpretation of cointegration analysis in this situation becomes more difficult. The shortcoming that is central to this thesis is of another nature. Cointegration has fairly strict data requirements and this can be very limiting to researchers when faced with less than perfect data. The imperfections can be missing observations, changing data frequencies, measurement error, overly simplistic data interpolation, etc. Many methods exist for interpolation of data to deal with the problem of missing observations, such as data imputation, the Parzen estimator and the Chow-Lin method (Datta & Du 2012). However, interpolation will always be a second best solution and it might be more productive to research new empirical approaches that are more robust to the data imperfections inherent in empirical data, such a cluster analysis.

2.4 Recent literature

This section highlights four recent articles that use the cointegration method to analyse market integration. Since the data used in the analysis section of this thesis is based on the world wheat price, the selected articles all focus on this particular commodity.

Goychuk & Meyers (2014) research the short and long run wheat price dynamics between Russia and Ukraine, countries that have become important players in the international wheat market, and the United States, the European Union and Canada from July 2004 to October 2010. They find that Russian wheat prices are not cointegrated with Canadian prices, but do find cointegration between prices in Russia, the United States and the European Union. The difference in quality between Russian and Canadian wheat is offered as an explanation for the lack of cointegration between these two markets. Ukrainian wheat prices are found to only be cointegrated with prices in the European Union. The results thus show that the Black Sea grain market is integrated to a degree with the international market. Long run price transmission elasticities indicate that price signals are transmitted and this enables the efficient allocation of resources. Furthermore, the authors find that price shocks are transmitted to the Russian and Ukrainian wheat markets symmetrically.

Wheat is the main staple crop in Pakistan, but the country is unable to produce a sufficient amount internally. Hence, local production is supplemented with imports to meet local demand and ensure food security. Mukhtar & Ishfaq (2009) study the impact of the Uruguay Round Agreement on Agriculture for Pakistan's wheat prices. The expected impact of this particular agreement is a 7 to 11 percent increase in global wheat prices. If the Pakistani wheat market is not cointegrated with world

prices, however, the implementation of the agreement has no implications for Pakistan. Cointegration reveals a stable long run relationship between international wheat prices and local wheat prices in Pakistan. This implies that the agreement will have consequences for the local economy and the authors analyse the specific welfare impact in the remainder of the article, providing a number of recommendations to deal with the negative effect on poor consumers.

Bakucs et al. (2012) analyse the price transmission of wheat between Germany and Hungary using weekly prices between January 2003 and September 2007. The authors use a Markov-Switching Vector Error Correction model, which allows time series analysis under different regimes. They find a cointegrating relationship and use three different regimes to fully capture the dynamics in the price transmission relationship. Interestingly, the most compelling regime, which has the highest probability and captures the largest share of the data, finds that the law of one price does not hold between Germany and Hungary.

Ghosh (2003) investigates the degree of market integration between regional wheat markets in India using monthly data between March 1984 and April 1997. Regional markets should be spatially integrated for consumers and producers to realize the gains from trade liberalisation. Stable long run relationships are found between these markets, despite their geographical dispersion. This implies that price signals are correctly transmitted within and between regions, which is a significant result for market liberalisation and agricultural price policy. A limited degree of intervention will allow private traders to contribute in the integrated markets, ensuring food and price stability at minimal expense for the Indian government.

2.5 Conclusion

Market integration is a complex, multidimensional concept, despite being seemingly straightforward at first glance. It is commonly measured by the degree of price transmission between spatially distinct regions in response to shocks. When considering market integration it is important to distinguish it from competitive equilibrium. The distinction is important if one is to draw correct conclusions from empirical data on prices and trade volumes. This chapter also identifies six factors that reduce the degree of market integration, namely commodity heterogeneity, market power, imperfect information, the temporal character of trade, trade policy and transaction costs. Next, several empirical methods were discussed that are used in econometric time series analysis to determine the degree of market integration between markets. The cointegration approach is the most commonly used method and is an important analytical tool that is able to identify the degree of integration, adjustment speed and asymmetry in price relationships. However, non-stationary transaction costs and sub-par data are problematic when applying cointegration. The importance of the cointegration approach is illustrated by a brief discussion of four recent articles that analyse market integration in different wheat markets.

3. Methodology

This chapter highlights the two methodological approaches applied in this thesis to investigate market integration. The chapter begins with an overview of cointegration, then discusses the different aspects of cluster analysis and finishes with a brief discussion of both methods with regard to market integration.

3.1 Cointegration

As discussed in chapter 2, cointegration has become the primary tool used by economists to investigate market integration in recent years. This section presents a concise discussion on non-stationary time series and spurious regressions, while also discussing the background of cointegration and vector error correction models (VECMs).

3.1.1 Non-stationary time series and spurious regressions

Non-stationarity and spurious regressions lay the foundation for understanding cointegration. This section briefly reviews and explains these concepts. Dougherty (2011) offers an excellent explanation of the difference between stationary and non-stationary time series and its implications. A time series of a commodity price such as wheat can be considered strictly stationary if the entire distribution is independent of time. However, economists are typically only concerned with the mean, variance and covariances of time series. Therefore it is sufficient to only impose that these three moments are independent of time (Verbeek 2012). This is referred to as weak stationarity or covariance stationarity. In general, a time series is said to be weakly stationary when its distribution satisfies the following three conditions:

1. Mean is constant and independent of time. ($E(X_t) = \mu$ for all t)
2. Variance is constant and independent of time. ($var(X_t) = \gamma_0$ for all t)
3. Covariances are constant and independent of time. ($cov(X_t, X_{t-k}) = \gamma_k$ for all t and k)

When a commodity price time series violates any of the above three conditions it is said to be non-stationary. Three common examples of non-stationary series are a random walk, a random walk with drift and a deterministic trend (Dougherty 2011). A random walk violates the second condition of weak stationarity, causing variance to increase over time. When a process is described by a random walk with drift, it fails to satisfy the first two conditions. Hence, in this case both the mean and variance of the series change over time. Finally, a deterministic trend does not comply with the first condition specified above, causing the mean to change over time.

If a non-stationary commodity price time series can be made stationary by differencing, it is considered to be difference-stationary. Differencing simply implies that the price changes of the series from one period to the next are taken as the new values. Hence, the first difference (Δ) of X_t is equal to $X_t - X_{t-1}$. A series is integrated of order 1, or I(1), if it can be transformed into a stationary series by differencing once. When the series has to be differenced twice to become stationary it is I(2), and so on. A stationary series is thus I(0), since it is already stationary and no differencing is required. A non-stationary commodity price time series can also be trend-stationary, which is the case if the series becomes stationary after extracting the time trend. (ibid)

The problem with non-stationary commodity price series is that they lead to two types of spurious regressions. First, consider two price series of markets that are in autarky (e.g. X_t and Y_t) and are both characterized by a deterministic trend. If one would run a regression of Y_t on X_t, the common dependence on time would make it seem like these prices are correlated, provided the sample is large enough. However, since the markets are operating completely independent, no real causal relationship exists. The second type of spurious regression occurs with price series that are considered to be random walks. Consider again a regression of Y_t on X_t, but this time these two variables are generated as independent random walks. Since the prices are again unrelated, the expected value of the slope-coefficient β_2 is zero. This is not the case, however, which implies that the estimator is inconsistent. The immediate result of both types of spurious regression is that there is a considerable risk of obtaining significant regression results, while no real relationship exists between the markets under investigation. Under these conditions regression results will typically be characterized by a rather high R^2, a high degree of autocorrelation between residuals and a statistically significant slope coefficient (β_2) estimate (Verbeek 2012). Hence, ignoring the potential for spurious regressions with non-stationary time series analysis can lead researchers to draw erroneous conclusions about market integration.

Since non-stationarity is highly problematic, econometricians have developed the unit root test to detect it in time series. The most common way of testing for a unit root is the augmented Dickey-Fuller (ADF) test, which uses adjusted values for the critical t-values. The general form of this test for an AR(p) process is as follows (ibid):

$$\Delta X_t = \beta_1 + \left(\beta_2 + \cdots + \beta_{p+1} - 1\right)X_{t-1} + \beta_3^*\Delta X_{t-1} + \cdots + \beta_{p+1}^*\Delta X_{t-p+1} + \varepsilon_t \qquad (3.1)$$

where β_3^*,...,β_{p+1}^* are suitably chosen constants. This test has the null hypothesis that $\left(\beta_2 + \cdots + \beta_{p+1} - 1\right)$ is equal to zero, implying that the series is non-stationary. Hence, this null hypothesis needs to be rejected for a time series to be considered stationary. However, since unit root tests often have low power (ibid), there is a significant chance of making a Type II error and fail to reject the null hypothesis when it is in fact false. Therefore, it is useful to also apply a statistical test to the data which uses the null hypothesis of stationarity. In this thesis the Kwiatkowski, Phillips, Schmidt and Shin (KPSS) test is used for exactly this purpose. The null hypothesis of stationarity implies that the random walk component must have a variance of zero, which is tested with a Lagrange multiplier test.

This thesis focusses on market integration between spatially distinct markets, using prices as the key determinant. However, as explained in this section, economic time series are often integrated of some particular order. This implies that a standard regression will not provide any useable information, due to the spurious nature of regressions featuring non-stationary time series. Fackler & Goodwin (2001) explain that commodity price data is likely to be non-stationary, due to common factors such as inflation, population growth or climate related shocks. In this case a commodity price in two distinct markets may seem to be moving jointly, while no real market integration is present. Hence, regular regression analysis is not an appropriate method to investigate market integration.

3.1.2 Cointegration

When applying OLS, one is investigating the conditional correlation of two or more time series. Generally speaking, the linear combination will be integrated to the same degree as the most highly integrated individual series (Dougherty 2011). For example, a linear combination of an I(2) and an I(1) commodity price time series will be I(2). However, this result does not hold if the price series share a long-run relationship and are integrated of the same order. Consider two I(1) price series for wheat, Y_t and X_t, that share such a long-run relationship. In this case, there will exist a value for β such that the linear combination $Y_t - \beta X_t$ is I(0), with $(1, -\beta)'$ generally being labelled as the cointegrating vector. This implies that the difference between the two series is stationary and relatively stable. In other words, the two series will wander together instead of drifting apart. When this theoretical relationship is present, the series are said to be cointegrated and will share a common stochastic trend (Verbeek 2012). This co-movement of prices is consistent with the law of one price discussed in chapter 2 and it can be concluded that these markets are integrated to a degree. When more than two variables are included in the model, there could be multiple cointegrating relationships. If the model includes k variables, the theoretical maximum to the number of cointegrating relationships is $k - 1$ (Dougherty 2011). Hence, this method allows the researcher to conduct a multivariate analysis of market integration between groups of markets.

In the case of cointegration, the problem of spurious regressions disappears and no longer invalidates the regression results, implying that the β coefficient is consistently estimated by the OLS procedure. The estimator is, in fact, considered to be super consistent because it converges to the true value of β at a much faster rate than usual. Verbeek (2012) provides an intuitive explanation for this unusual result. Provided that the price series under investigation are cointegrated, the residuals of the regression are I(0) for the true value of β and must be stationary. For any parameter estimate that is not close to this true value, the residuals will be non-stationary and have a large variance. However, when the estimate converges to the true value, this variance will be much smaller. Since OLS is chiefly concerned with minimizing the variance of the residuals in the sample, it is very capable of finding estimates close to the true value of β.

It is thus crucially important to distinguish between cases of cointegration and spurious regression when investigating market integration based on price series. As mentioned above, a straightforward way of testing for cointegration is to test whether the residuals are I(0). The ADF test discussed in the previous sub-section can be applied to test this. However, there is a complication when testing for unit roots in OLS residuals. Since the OLS estimator minimizes the variance of the residuals, it makes the residuals look as stationary as possible, even in the absence of a cointegrating relationship. For the ADF test to provide valid results, the critical values have to be adjusted downward. When testing for cointegration between more than 2 variables, the appropriate critical values become even more negative. Appropriate values for the test can be found in standard econometric textbooks (e.g. Verbeek 2012, p. 349). A procedure for testing the number of cointegrating relationships between multiple series is the Johansen trace or maximum eigenvalue tests (ibid). In the case of multiple cointegrating vectors, they are jointly referred to as the cointegrating space. The total number of independent cointegrating vectors is $r \leq k - 1$, with r being the rank of the matrix β (also known as the cointegrating rank of $\vec{Y}_t$) and k the number of price series under consideration.

3.1.3 Vector Error Correction Models

As shown in the previous sub-section, it is possible to identify a long-run relationship between two or more price series of a commodity in the case of cointegration, indicating market integration. However, a cointegrating relationship doesn't provide any details on the short-run dynamics of the relationship. When a set of two or more prices are cointegrated, it is possible to depict the data with a vector error-correction model (VECM) that captures both short and long-run dynamics between the markets (Dougherty 2011). Consider the following general VAR(p) model for k price series:

$$\Theta(L)\vec{Y}_t = \delta + \vec{\varepsilon}_t \tag{3.2}$$

where $\Theta(L)$ is the $k \times k$ matrix lag polynomial and each element represents a p^{th} order polynomial in L. This can be written as a VECM, resulting in the following equation (Verbeek 2012):

$$\Delta\vec{Y}_t = \delta + \Gamma_1\Delta\vec{Y}_{t-1} + \cdots + \Gamma_{p-1}\Delta\vec{Y}_{t-p+1} + \Pi\vec{Y}_{t-1} + \vec{\varepsilon}_t \tag{3.3}$$

where

$$\Gamma_i = -\sum_{j=i+1}^{p} \Theta_j \tag{3.4}$$

$$\Pi = \sum_{i=1}^{p} \Theta_i - I_k \tag{3.5}$$

The long run matrix described by equation 3.5 determines the long-run dynamics of $\vec{Y}_t$ and it contains the error correction terms. The rank of this matrix is equivalent to the number of cointegrating relationships between the different markets. Three distinct possibilities arise. First, if the matrix has a rank of zero, no cointegrating vectors exist. Second, if all variables in $\vec{Y}_t$ are I(0), the matrix must be of full rank (k) and there is no need to estimate the VECM (a VAR representation is sufficient). Third, when the rank of the matrix is equal to $(0 < r < k)$, there are r cointegrating vectors representing stable long-run relationships between the markets under consideration. In this case it is possible to write the Π matrix as the product of a $k \times r$ matrix γ and an $r \times k$ matrix β'. Hence, we end up with the term $\gamma\beta'\vec{Y}_{t-1}$ in equation 3.3. Here $\beta'\vec{Y}_{t-1}$ represents the cointegrating space with r cointegrating relationships and γ measures the speed of adjustment of the price series contained in $\vec{Y}_t$. The VECM thus enables us to conduct a multivariate analysis of market integration, providing insight in short and long-term dynamics of the market as a whole. The model results in parameter estimates for the short run response of a price in a particular market to price changes in other markets. Moreover, the cointegrating relationship reveals the long run response to price changes in other markets and the adjustment speed parameter provides insight into the temporal character of price adjustment with market integration.

3.2 Cluster analysis

This section serves to briefly introduce the cluster analysis method. Cluster analysis is a so called interdependence technique, which allows researchers to identify underlying structures among sets of variables (Hair et al. 1998). It defines groups with the highest degree of homogeneity, while ensuring that there is maximum heterogeneity between the different groups. This section focusses on hierarchical agglomerative methods, where the k cluster solution is created by combining two clusters from the $k+1$ cluster solution (Lattin et al. 2003). Furthermore, the primary objective is to identify relationships among the different time series as an indication of market integration, based on a set of non-parametric statistics that characterizes each individual series. The disadvantage for cluster analysis is its subjective nature and the fundamental role of the researcher's judgement in selecting the appropriate similarity metric and linkage criterion (Hair et al. 1998). The different choices available with respect to the metrics and linkage criteria are discussed first. Next, stopping rules and the importance of selecting appropriate variables of interest that characterize the different time series are discussed.

3.2.1 Similarity metric

Before getting into the specifics of the different similarity metrics available, it is important to explicitly state several underlying assumptions (Gordon 1999). First, the focus is on pairwise distances between objects. Second, the following three conditions for distances between objects have to be met:

$$d_{ij} \geq 0 \tag{3.6}$$

$$d_{ii} = 0 \tag{3.7}$$

$$d_{ij} = d_{ji} \tag{3.8}$$

where d_{ij} represents the distance between objects i and j. These conditions rule out asymmetric distances between objects and imply that the matrix of similarities is fully specified by the $n(n-1)/2$ values in its lower left triangle. Finally, the similarity matrix is assumed to be metric and Euclidean. It is metric if it satisfies the triangle inequality ($d_{ij} \leq d_{ik} + d_{kj}$) for all possible trios of objects i, j and k. The Euclidean assumption implies that there is a selection of points (P) in Euclidean space where the distance between P_i and P_j is equal to d_{ij}.

The most common distance metric applied to variables measured on ratio or interval scales is the Euclidean distance. This metric is a special case of a general class of distance metrics known as the Minkowski p-metrics (or L_p metrics), which are calculated as follows (Lattin et al. 2003):

$$d_{ij}(p) = \left[\sum_{k=1}^{v} \left(x_{ik} - x_{jk}\right)^p\right]^{\frac{1}{p}} \tag{3.9}$$

where x_{ik} represents the value of the k^{th} variable for object i and the total number of variables is equal to v. The Euclidean distance is equal to the L_2 metric. To ensure that each variable has equal weight in the final outcome, the data is typically standardized. Hair et al. (1998) provide an example

to illustrate the importance of this point, which is adapted to fit the context of this thesis. When attempting to cluster the time series based on a number of variables that characterize the series, the variables with a larger degree of variation generally have more impact on the final similarity metric calculated. The mean of a series, for example, can have a variance several orders of magnitude larger than the variance of the skewness or kurtosis of a series. This is why standardizing variables is often a sensible step to take.

In equation 3.9 the value for p is restricted to a range of 1 to ∞, however a value of 2 seems most appropriate for measuring the direct distance between standardized variables. Other values for L_p are only appropriate in special cases (Lattin et al. 2003). For example, L_1 is known as the city block metric, where the distance between points i and j is comparable to walking along a grid of parallel and perpendicular streets. A variation on the Euclidean distance is the squared Euclidean distance, which simply involves squaring equation 3.9. The advantage of this variation is the fact that the square root no longer has to be computed, which significantly speeds up the procedure (Hair et al. 1998).

An alternative to the Minkowski p-metrics is the Mahalanobis distance, which takes into account covariance in the data. This approach performs standardization on the data and also adjusts for intercorrelations among the variables (Hair et al. 1998). It is, however, unlikely that the nonparametric statistics used to characterize the time series in the cluster analysis will be affected by this to a high degree. Moreover, this particular distance metric is not available in STATA by default and will therefore not be considered any further.

The Canberra similarity metric is suitable for comparing the distance between objects i and j (Gordon 1999). This metric returns values between 0 and v, the number of variables. It is particularly sensitive to minor changes when x_{ik} and x_{jk} are close to zero, but is less reliable when the variables are sample estimates of some quantities (ibid). It is a generalization of metrics based on the binary presence or absence of variables.

Finally, there are two similarity metrics that are primarily concerned with the relative magnitudes of the different variables (ibid). The characterizing variables of a time series form a vector with p components and attention is restricted to comparing the direction of the vector of each series. These correlation-type measures are known as the angular separation and the correlation coefficient. In essence they measure the cosine of the angle between two vectors. In angular separation the vector is measured from the origin, while for the correlation coefficient it is measured from the mean of the data. Values for both metrics range from -1 to 1. Since these two metrics are primarily concerned with relative magnitudes (Hair et al. 1998) they are not particularly appropriate for the cluster analysis involving time series data. As mentioned in the previous chapter, time series may seemingly develop in a similar manner, resulting in a high degree of correlation, due to an underlying time trend such as inflation. However, this does not imply that the markets are integrated to any substantial degree. Hence, using distance measures rather than correlational measures is more appropriate when considering the goal of this thesis.

Hair et al. (1998) recommend using several similarity metrics and comparing the results with theoretical patterns expected by the researcher. This is a sensible procedure primarily due to the fact that different metrics can result in different cluster solutions.

3.2.2 Linkage criteria

There are a plethora of different algorithms available that can identify similarities between the commodity price time series and assign them to clusters. The primary criterion of all these different algorithms is to maximize the difference between clusters, while minimizing the within cluster difference (Hair et al. 1999). Most of these approaches can be classified in one of two categories, namely hierarchical and non-hierarchical. Non-hierarchical clustering, which generates a predefined number of clusters and requires a specific seed point, is not considered here. Hierarchical clustering creates a tree like structure, where an object cannot be assigned to more than one cluster and all objects are assigned to one of the clusters available (Lattin et al. 2003). Furthermore, the specific focus is on agglomerative clustering. With this method each object starts out in its own cluster and, through subsequent steps, clusters are joined together until only one remains, which then contains all objects. The final result of this procedure is depicted graphically in the form of a dendrogram. Five of the most commonly used agglomerative algorithms, which differ in how they calculate distance between clusters, are concisely discussed below.

Single linkage is based on minimum distance between clusters (Hair et al. 1998). The two time series, as described by a common set of characterizing non-parametric statistics, that are most similar to each other are jointly placed in the first cluster. Then the next shortest distance is calculated and either a third series joins the existing cluster or a new two member cluster is formed. This procedure continues until only one cluster remains. The main drawback of this approach is that a series is added to a cluster, so long as it is close to one of the series in the cluster, despite the fact that it might be very dissimilar to other series in the cluster. The direct result is long, stringy clusters and potential non-convexity of clusters (Lattin et al. 2003).

Complete linkage works in a similar manner to single linkage, except for the fact that in this procedure the distance between the farthest pair of series is used (ibid). In this manner one ensures that a series that is added to a cluster is close to all series in the cluster and not just one (as with single linkage). The maximum distance between the members of two clusters denotes the minimum diameter of a sphere that encloses objects in both clusters (Hair et al. 1999). This procedure eliminates the stringy cluster problem of single linkage and is also more likely to generate convex clusters. However, it can be highly sensitive to outliers (Lattin et al. 2003).

The average linkage procedure, perhaps unsurprisingly, operates in a similar manner as the previous two, but uses the average distance between all series in one cluster and all series in another cluster. It can therefore be considered a sort of compromise between single linkage and complete linkage. The procedure is less sensitive to extreme outliers and uses information from all cluster objects instead of only one (Hair et al. 1999). The resulting clusters typically have small within cluster variation and the method is also biased in the production of clusters with similar variance. Instead of the average one can also use the median, which is known as median linkage (Lattin et al. 2003).

The centroid linkage method is also based on cluster averages. However, the key difference with average linkage is the fact that it is not based on the distance between all pairs of time series, but instead on the average (or centroid) of the clusters itself (ibid). Hence, every time a new series is joined with an existing cluster, its centroid shifts. This can produce conflicting results, as reversals may occur (Hair et al. 1998). This method is quite robust to outliers, but might be outperformed by average linkage (Lattin et al. 2003).

Ward's method uses a slightly different approach than the previous four methods, as it focusses on agglomerating clusters that result in the smallest within cluster sum of squares (ibid). Hence, distance between clusters is measured by the sum of squares between them summed over all time series (Hair et al. 1998). The procedure tends to result in clusters with a similar number of objects and often merges clusters with a small number of observations.

3.2.3 Stopping rules

One of the fundamental difficulties after conducting a cluster analysis is assessing how many clusters are actually present among the different time series. Lattin et al. (2003) point out that for agglomerative clustering there is no definitive answer to this question. The output of this type of analysis produces a dendrogram, which contains everything from the n-cluster solution up to the one cluster solution. The authors recommend looking for a relatively wide distance over which the number of clusters does not change. However, it is also argued that analysing a dendrogram involves a substantial amount of subjectivity and relies to a large extent on the researcher's judgement. Hair et al. (1998) support this notion and claim that there is no standard, objective selection procedure. Their recommendation involves analysing a number of different cluster solutions and deciding among them based on common sense and theoretical foundations.

Many different stopping rules have been developed over the years by researchers, which all rely on different criteria. STATA offers two stopping rules by default, a global and a local one. Global stopping rules attempt to find the optimal number of clusters based on all available data, while local stopping rules focus on the examination of pairs of cluster and whether or not they should be combined (Gordon 1999). The global rule is the Calinski and Harabasz index, which analyses the ratio between the total between-cluster sum of squared distances and the total within-cluster sum of squared distances. The local rule is the Duda and Hart index; it is a decision rule for combining two subclusters based on the within-cluster sum of squared distances. For both indexes larger values indicate more distinct clusters. Milligan & Cooper (1985) evaluated 30 stopping rules and found that the aforementioned two rules performed the best.

3.2.4 Variable selection

The objective of cluster analysis in this thesis is to identify relationships among different sets of time series from spatially distinct markets. A close relationship between two series can be interpreted as an indication of market integration between the pair of markets. The selection of variables that characterize the different time series is thus a critical step, as the result of the cluster analysis is constrained by the variables which are selected by the researcher (Hair et al. 1998). Cluster analysis reveals the inherent structure in the data based on these variables and it cannot distinguish between relevant and irrelevant variables. Hence, variable selection is vital and only those variables that describe the evolution of the different series and differ between the series to a significant degree should be included. Furthermore, as pointed out in sub-section 3.2.1, the variables should be standardized to avoid variables with substantial degrees of variance from dominating the final outcome (Gordon 1999).

3.3 Discussion

Cointegration focusses on the difference between series; if a cointegrating relationship exists these differences should be stationary. This is consistent with the law of one price, where the difference in price between two markets is determined by the transaction costs. Hence, cointegration is a straightforward method that enables researchers to review market integration. The technique reveals long-run relationships between integrated variables and this is a useful starting point for investigating market integration. If a cointegrating relationship is found between the prices of a specific commodity in different markets, price signals will be transmitted between these markets to a certain extent in a long-run equilibrium. However, simply estimating the long-run relationship does not provide all the information in the case of market integration. Rapsomanikis et al. (2006) further label the cointegrating parameter an atheoretical statistical concept that may not have an economic interpretation similar to the parameters in a structural model.

When applying the VECM, a distinction between short and long-run effects can be identified, as well as the adjustment speed. These different aspects of market integration are assigned very specific parameter values with this type of analysis. Rapsomanikis et al. (2006) argue that it is perhaps the most useful tool for market integration analysis, since it allows for the gradual rather than instantaneous adjustment of prices. However, when considering data that has serious shortcomings, these exact parameter estimates quickly lose their value. In this case the technique may even give the researcher a false sense of exactness, which is clearly not supported by the data. Additionally, if one aims to investigate market integration between large numbers of markets, the VECM becomes increasingly difficult to interpret. Despite these shortcomings the cointegration approach is currently one of the most applied methods in market integration research and for this reason it is used as a benchmark for cluster analysis.

Cluster analysis based on characterizing variables of the different time series can provide a valuable alternative to the cointegration approach. It is less sensitive to data imperfections and provides its output in the form of a graphical representation that is fairly simple to interpret. Due to this, the method should also make market integration research among a large number of markets easier to grasp for non-econometricians. The disadvantage of cluster analysis is the subjective nature of the analysis and the lack of any statistical tests that can distinguish between significant and insignificant findings. However, when considering the data quality that is available from developing countries, this method may be the best approach available when analysing market integration. In short, if the underlying data is shoddy, VECMs provide unreliable estimates and one can only make general statements about the degree of market integration. Hence, cluster analysis might prove to be a worthwhile alternative.

4. Data

This chapter introduces the data, and explains how and why the series that are ultimately used were constructed. The analysis uses artificial rather than real world data, which is sensible when considering the overall goal of this thesis, and it will be briefly discussed how this data was generated.

4.1 Original series

As stated in the previous chapters, the primary goal of this thesis is to compare the cointegration approach with cluster analysis. In order to do this in a satisfactory manner one needs to have a complete dataset with a reasonable amount of distinct series, which will allow for experimentation. Unfortunately such a dataset does not exist. The FAO GIEWS price tool (FAO 2015) is the most likely candidate, as it features time series data on several commodities across a large number of countries. However, the dataset is lacking, featuring a substantial amount of missing data, and therefore it cannot fully provide the required information for an in depth comparison between the two methods. Since no appropriate dataset is available, a number of time series are artificially generated. To ensure these artificial series represent real world data as closely as possible they were generated based on monthly data from the World Bank GEM commodity database (World Bank 2015). The selected commodity is hard red winter wheat (US) and prices are in nominal US dollars per ton (see figure 4.1).

Figure 4.1: World Bank GEM Commodity database (2015) - Wheat price

4.2 Data generation

Based on the original series specified above, a total of eleven series have been generated (see figure 4.2). The rationale behind this is to generate a number of different market conditions, representing potential conditions within countries. A number of these series are designed to be strongly integrated with the original series, while others feature a minor degree of integration or even move independently from the original series. By generating a number of these series, with known properties, a direct comparison between them is straightforward. Comparing multiple markets simultaneously is one of the expected strengths of cluster analysis, but to ensure a comparison with the cointegration approach is possible the number of markets has to be kept reasonably low. Since

correctly interpreting a vector error correction model consisting of a very high number of markets is challenging, a total of twelve series provides a suitable compromise.

Prices in the first two markets are designed to develop completely independent from the original series. They differ among themselves in their slope and the variance of the stochastic error term (unrelated market 1 $\varepsilon_t \sim N(0,4)$; unrelated market 2 $\varepsilon_t \sim N(0,8)$). The exact formulas and parameters used for these two series are specified in table 4.1, where z_t represents the original series and y_t is the generated series under consideration. These two series have been added in order to test the accuracy of cluster analysis; since the two series are unrelated to the original series the method should be able to distinguish them as separate from the others.

Table 4.1: Unrelated markets specification (Random walks with drift)

Series	ID	Initial value	Equation	$\boldsymbol{\beta_1}$
Unrelated market 1	1	$y_1 = z_1 + x;$ $x \sim N(0,10)$	$y_t = \beta_1 + y_{t-1} + \varepsilon_t$	0.4
Unrelated market 2	2	$y_1 = z_1 + x;$ $x \sim N(0,10)$	$y_t = \beta_1 + y_{t-1} + \varepsilon_t$	0.6

Trade policy measures are often undertaken by governments to protect specific groups within society (Krugman et al. 2012). Typically the group that is supported has a certain amount of influence on the political process and despite the claim that national welfare is being protected, often the benefits accrue to individuals rather than society as a whole. Bellemare et al. (2013) investigate the case of commodity price stabilization through government intervention and conclude that, contrary to common wisdom, welfare gains from a reduction in price volatility are regressive in nature. The implication being that wealthier households gain more than poorer households from this kind of intervention. Regardless of the actual impact on national welfare, the fact remains that governments in developing countries often choose to intervene in commodity markets in an attempt to stabilise prices (Timmer 1989; Gouel 2013). A well-known example is Egypt, where food subsidies have been in place in one form or another since the 1940s (Löfgren & El-Said 2001). In order to replicate this behaviour two series have been added that represent a, admittedly oversimplified, version of this particular kind of behaviour. In both series prices are set by the government in relation to the world market price based on a set of decision rules (see table 4.2).

Table 4.2: Fixed regime specification

Series	ID	Decision rules
Fixed regime 1	3	If original series < 200, value is 100; If original series ≥ 200 and < 300, value is 180; If original series ≥ 300 and < 400, value is 250; If original series ≥ 400, value is 300
Fixed regime 2	4	If original series < 160, value is 80; If original series ≥ 160 and < 240, value is 140; If original series ≥ 240 and < 300, value is 200; If original series ≥ 300 and < 400, value is 250; If original series ≥ 400, value is 280

The remaining seven series are all integrated with the original series to a certain degree and are described by equation 4.1, which is a variation of the error correction mechanism (Rapsomanikis et al. 2006; Verbeek 2012). This particular kind of model features short term price adjustments in the form of lagged differences and also includes a long term relationship (the cointegrating relationship) that moves prices of the generated series into long term equilibrium with the original series. The amount of lagged differences included determines how long a short term shock influences the development of prices and this has been restricted to two in order to avoid overly complex lag structures.

$$\Delta y_t = \beta_1 \Delta z_{t-1} + \beta_2 \Delta z_{t-2} + \beta_3 \Delta y_{t-1} + \beta_4 \Delta y_{t-2} - \beta_5 (y_{t-1} - \beta_6 z_{t-1}) + \varepsilon_t \qquad (4.1)$$

In the above equation the stochastic term is defined as $\varepsilon_t \sim N(0,2)$ for all integrated series. The degree of integration ranges from very strong to very weak, to simulate different market conditions. The most strongly integrated series has a high adjustment speed parameter β_5 and its price level strongly depends on the lagged differences of the original series (β_1 and β_2). The very weakly integrated series, on the other hand, has a low speed of adjustment and its price level depends strongly upon its own lagged differences (β_3 and β_4) rather than the lagged differences of the original series. An example of a highly integrated market is the United States and a weakly integrated market can be characterized by a landlocked country with poor physical infrastructure. The parameter information is given in table 4.3. Series 10 and 11 have the same properties as the medium series; however both feature deviations from the cointegrating relationship. Series 10 has a government imposed price ceiling of 280 USD/ton, whereas series 11 has a simulated crop failure in September of 2003 and a price ceiling of 300 USD/ton from February 2008 to September 2008. It should be noted that the exact values of the coefficients have been chosen in an arbitrary fashion and one could easily have used different numbers, provided the relative strength of the relationships is maintained. Furthermore, the βs of all integrated series fluctuate, due to a random addition of $N(0; 0.001)$.

Table 4.3: Cointegrated series specification

Series	ID	Initial value	β_1	β_2	β_3	β_4	β_5	β_6
vStrong	5	$y_1 = z_1 + x$; $x \sim N(0,2)$	0.3	0.03	0.01	0	0.5	1
Strong	6	$y_1 = z_1 + x$; $x \sim N(0,2)$	0.2	0.02	0.05	0	0.25	1
Medium	7	$y_1 = z_1 + x$; $x \sim N(0,2)$	0.1	0.01	0.1	0.01	0.125	1
Weak	8	$y_1 = z_1 + x$; $x \sim N(0,2)$	0.05	0	0.2	0.02	0.0625	1
vWeak	9	$y_1 = z_1 + x$; $x \sim N(0,2)$	0.01	0	0.3	0.03	0.03125	1
Medium with ceiling	10	$y_1 = z_1 + x$; $x \sim N(0,2)$	0.1	0.01	0.1	0.01	0.125	1
Medium with shock and intervention	11	$y_1 = z_1 + x$; $x \sim N(0,2)$	0.1	0.01	0.1	0.01	0.125	1

The eleven generated series are used to conduct the analyses in the next chapter. The monthly data from January 1998 to October 2014 are used, resulting in a total of 202 observations for each series. The series will be analysed using their IDs as specified in the above tables, rather than their full names, for the sake of brevity. The artificially generated data enables a comprehensive analysis of the two methods under consideration. Moreover, since the data has no missing observations it is possible to conduct a sensitivity analysis and investigate the exact effect of missing data on the final results of the cluster analysis. This would not have been possible using real world data due to the fact that finding a suitable number of perfectly continuous series is difficult, if not impossible. The alternative approach of data generation is thus a necessary step and in line with the overall goal of the thesis.

Table 4.4 shows the correlation matrix of all twelve series. It reveals that all series under consideration have a strong degree of correlation, with the minimum value being 0.525 between series 2 and 3. However, as discussed in the previous chapters, this correlation is by no means an indication of a long run relationship between the different series. The series can simply share the same trend and merely appear to be moving jointly. For a true long run relationship to exist further evidence is required such as the existence of a cointegrating relationship, which is discussed in the next chapter. The table does confirm that series 5 is strongly correlated with the original series, as it should be by design. Furthermore, the two fixed regime series (3 and 4) are also closely correlated with the original series. This can be explained by the fact that the decision rules are specified in such a way that these two series respond instantaneously after a predefined price threshold has been crossed.

Table 4.4: Correlation matrix

ID	Original	1	2	3	4	5	6	7	8	9	10	11
Original	1.000											
1	0.853	1.000										
2	0.627	0.740	1.000									
3	0.946	0.797	0.525	1.000								
4	0.972	0.858	0.597	0.928	1.000							
5	0.977	0.864	0.654	0.927	0.950	1.000						
6	0.958	0.885	0.689	0.914	0.931	0.990	1.000					
7	0.925	0.914	0.740	0.880	0.901	0.960	0.986	1.000				
8	0.888	0.925	0.747	0.855	0.872	0.921	0.958	0.988	1.000			
9	0.838	0.938	0.731	0.814	0.839	0.867	0.906	0.951	0.980	1.000		
10	0.936	0.909	0.736	0.884	0.912	0.957	0.975	0.984	0.973	0.941	1.000	
11	0.918	0.910	0.755	0.868	0.901	0.947	0.972	0.987	0.980	0.945	0.977	1.000

Based on the data generation procedure it is reasonable to expect series 1 to 4 to be most distinct from the original series, as these are the unrelated markets and the fixed regimes. For the remaining series it is more difficult to predict which ones will resemble the original series most. The very strongly and strongly integrated series should naturally follow its price development very closely. However, it is not clear a priori whether the medium series which feature some form of deviation (10 and 11) will resemble the original series more closely than the weakly and very weakly integrated series (8 and 9). The analyses in the next chapter will have to shed some light on this matter.

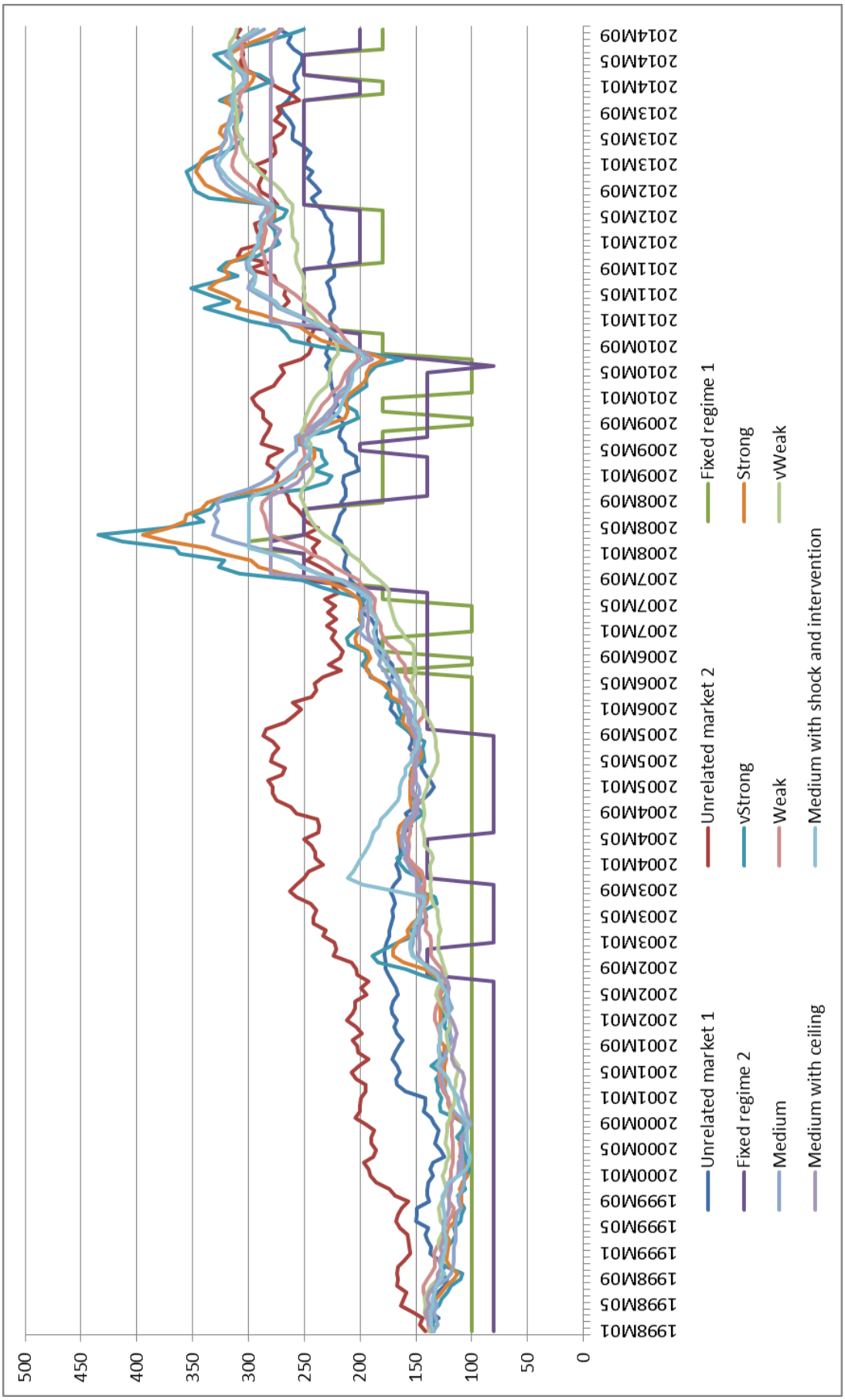


Figure 4.2: Generated time series

5. Analyses

The analyses of the data using both cointegration and cluster analysis are conducted in this chapter. Cointegration is discussed first and cluster analysis second. Moreover, a sensitivity analysis is performed to check the robustness of cluster analysis.

5.1 Cointegration

The first step in conducting the cointegration analysis is to check whether all series are integrated of the same order. The Augmented Dickey Fuller (ADF) and KPSS tests combined confirm that the original series and all eleven generated series are integrated of order 1. This implies that a cointegrating relationship between any of the series can potentially exist. Table 5.1 shows the results of the two analyses with * indicating significance at the 10% level, ** indicating significance at the 5% level and *** indicating significance at the 1% level. The reported KPSS test statistics are calculated with the quadratic spectral kernel in combination with the automatic bandwidth selection routine, as suggested by Hobijn et al. (1998). The KPSS test statistics should not be significantly different from 0, as this would indicate non-stationarity.

Table 5.1: Order of integration tests

Series	Order of integration	ADF Lags	ADF Test statistic	KPSS Test statistic
Original	1	6	-5.499 ***	0.047
1	1	4	-4.554 ***	0.044
2	1	10	-4.419 ***	0.044
3	1	9	-5.621 ***	0.031
4	1	12	-5.605 ***	0.028
5	1	6	-5.532 ***	0.049
6	1	6	-5.227 ***	0.064
7	1	6	-4.578 ***	0.090
8	1	6	-4.468 ***	0.101
9	1	0	-6.782 ***	0.103
10	1	0	-10.891 ***	0.078
11	1	0	-6.033 ***	0.063

Next, it is important to select the appropriate value for the number of lags to be included in the Johansen trace test and the final VECM. The lag order selection statistics can be used to determine the lag-order for a VECM with time series that are integrated of order 1 (Nielsen 2001). The four selection statistics are the final prediction error (FPE), Akaike's information criterion (AIC), Schwarz's Bayesian information criterion (SBIC), and the Hannan and Quinn information criterion (HQIC). In order to determine whether the generated series are cointegrated with the original series, as all but the unrelated markets should be by construction, bivariate Johansen trace tests are conducted. Each generated series is separately analysed in conjunction with the original series and the results of this analysis are displayed in table 5.2. The Johansen trace tests in this bivariate analysis confirm the a priori expectations and reveal that both unrelated markets (series 1 and 2) are not cointegrated with the original series, whereas all other series are.

Table 5.2: Bivariate lag length determination and Johansen trace tests (with the original series)

Series	1	2	3	4	5	6	7	8	9	10	11
Optimal lag length	2	2	2	4	3	2	2	2	2	2	3
Johansen trace test	0	0	1	1	1	1	1	1	1	1	1

Based on these results one can exclude the random walks with drift from further analysis in the bivariate framework. The remaining series are all placed in a bivariate VECM with the original series. For series 5 through 9 the analysis should reveal the cointegrating relationship and adjustment speed as specified in the previous chapter. For series 10 and 11 it is unlikely that the analysis will return the exact parameters specified in chapter 4, since a number of deviations from the relationship occur within each of these two series. Both series feature some degree of government intervention and series 11 additionally suffers a price shock in September of 2003. Series 3 and 4 are the fixed regime models, which follow the world price to an extent. The exact price set by the government depends directly on the world market price for both series. However, no cointegrating relationship was specified in the construction of both of these series and therefore the results of the analysis cannot directly be compared to the specification. Table 5.3 summarizes the results of the nine bivariate VECMs in a succinct way by providing the parameter estimates for the βs of the nine generated series with the standard errors reported in parentheses and the significance level indicated with the same notation as above. The specification of the different βs is the same as in chapter 4, i.e. β_1 and β_2 are the lagged differences of the original series, β_3 and β_4 are the lagged differences of the series under consideration and β_5 is the adjustment speed parameter. Additionally, the parameter estimates for β_6 represent the coefficient of z_{t-1} in the cointegrating relationship of equation 4.1. The number of lags used in the analysis is the same as specified in table 5.2. The parameter estimates for the original series are not reported, since our primary interest lies in the generated series and their associated parameter estimates. The nine adjustment speed parameters found in each of the separately fitted bivariate models should also not be significantly different from zero for the original series, as only the generated series are determined by this relationship while the original series is predetermined and fixed. This expectation holds for all but one of the bivariate VECMS; in the bivariate analysis between the original series and series 10 the adjustment speed parameter is significantly different from zero for the original series. This is likely due to the price ceiling imposed upon series 10, which affects 50 out of 202 periods.

Table 5.3: Parameter estimates for bivariate VECMs

Series	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$
3	0.181 (0.107) *	-	0.027 (0.086)	-	-0.481 (0.080) ***	-0.702 (0.033) ***
4	0.228 (0.113) **	-0.021 (0.108)	0.064 (0.113)	-0.142 (0.105)	-0.308 (0.099) ***	-0.790 (0.029) ***
5	0.234 (0.067) ***	0.080 (0.056)	-0.062 (0.060)	-0.012 (0.015)	-0.572 (0.067) ***	-1.002 (0.003) ***
6	0.172 (0.022) ***	-	0.008 (0.035)	-	-0.278 (0.021) ***	-1.003 (0.007) ***
7	0.100 (0.013) ***	-	0.074 (0.050)	-	-0.131 (0.011) ***	-1.001 (0.013) ***
8	0.030 (0.010) ***	-	0.188 (0.063) ***	-	-0.070 (0.007) ***	-0.950 (0.026) ***

9	0.011 (0.010)	-	0.317 (0.065) ***	-	-0.029 (0.004) ***	-1.075 (0.069) ***
10	0.133 (0.026) ***	-	0.099 (0.061)	-	-0.085 (0.018) ***	-0.858 (0.045) ***
11	0.112 (0.017) ***	-0.018 (0.019)	0.574 (0.068) ***	-0.171 (0.061) ***	-0.061 (0.012) ***	-0.956 (0.049) ***

For series 5 through 9 the parameter estimates for the cointegrating equation and the adjustment speed parameter (β_6 and β_5) are all reasonably close to the original specification. As the strength of the cointegrating relationship decreases, the accuracy of the β_6 parameter estimate does seem to decrease, as seen in series 9 where the value is -1.075 instead of -1. Furthermore, for series 10 and 11 the estimates for the cointegrating relationship and the adjustment speed parameters are both quite different from the original specification. As mentioned above, this was to be expected due to deviations from the relationship in both series. Finally, series 3 and 4 feature a fairly high speed of adjustment and a parameter estimate for β_6 smaller than -1. The high adjustment speed parameter estimate can be explained by the fact that both series respond instantaneously to changes in the world price when a predetermined threshold is crossed. Hence, a high adjustment speed is not unexpected in these equations. Furthermore, the fact that the parameter estimate for β_6 in the cointegrating equation is smaller than -1 can be explained by the decision rules as defined in chapter 4. Due to these rules series 3 and 4 are structurally below the original series, which translates to a parameter that is smaller than -1.

Compared to the original specification of chapter 4 the correct number of lagged difference terms is only found for series 5 and 11 by the optimal lag length test statistics. However, for the selected optimal lag lengths, the terms that are found to be significantly different from zero do have parameter estimates that are close to the expected values. The exception being series 11, which has parameter estimates for β_3 and β_4 that are significant at the 1% level but are nowhere near their original specification. One potential cause of this inability to correctly determine the number of lags and the associated parameters is the fact that in the original specification these parameters have rather small values, ranging from 0.3 to 0.01. However, for series 8 the fitted VECM does find a small, but nonetheless significant, parameter estimate of 0.03 for β_1, which is fairly close to the original specification of 0.05. Hence, other factors may be obfuscating the analysis here.

To analyse whether the parameter estimates for series 5 to 11 are in line with the original specification of chapter 4 or differ in a statistically significant manner, one can simply check if the true parameter value is contained within the parameter estimate's 95% confidence interval. This is not possible for series 3 and 4, since they do not have predefined parameter values. The confidence intervals for the parameter estimates of series 5 to 11 are given in table A.1 of the annex. When comparing the confidence intervals of the parameter estimates that were found to be significantly different from zero above with the true parameter values, one finds that for nearly all estimates the true value is within the 95% confidence interval. The only exceptions are β_5 and β_6 for series 10, and β_3, β_4 and β_5 for series 11. This clearly illustrates how deviations from the normal market relationship can significantly complicate the analysis and result in incorrect parameter estimates. Overall it can be concluded that the bivariate VECM analyses of the generated series versus the original series is by no means perfect, but nonetheless very capable of finding the correct cointegrating relationships and associated adjustment speed parameters in most cases.

To formally check whether the bivariate VECMs are correctly specified one can test the stability condition of the estimates, which is a way to check whether the number of cointegrating relationships has been correctly specified, and check for signs of autocorrelation in the residuals. The results of both tests can be found in table 5.4. For the stability test a result of 1 indicates that the number of cointegrating equations has been correctly specified, since we are conducting a bivariate analysis with only one cointegrating equation. Essentially, with this test one should find that with K variables and r cointegrating equations there are a total of $K - r$ unit eigenvalues. The autocorrelation test was conducted with 4 lags and if there is no sign of autocorrelation the specification is correct, whereas finding evidence for the presence of autocorrelation implies that the model is potentially incorrectly specified. The results confirm that the models have been correctly fitted. The result of the autocorrelation test for series 7 and 10 does indicate a potential problem, finding autocorrelation in the second and fourth lag respectively. This is not deemed significant enough to reject the specification altogether.

Table 5.4: Results of the stability condition and autocorrelation tests

Series	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Stability test	1	1	1	1	1	1	1	1	1
Autocorrelation	No	No	No	No	Yes	No	No	Yes	No

Another way of analysing the series is to jointly place them in a single multivariate VECM. This is particularly interesting when analysing a larger number of spatially distinct markets and one is not only interested in ascertaining how well integrated each individual market is with the world market, but also how these markets influence each other. An analysis of several markets in close proximity with each other could, for example, also reveal a cointegrating relationship between these markets rather than just individual long-run relationships with the world market. Not to mention the fact that in the real world a crop failure in one market is very likely to cause some sort of price response in other nearby markets. Furthermore, one of the biggest strengths of cluster analysis is analysing large groups of markets and placing them into closely related clusters. Hence the comparison between the two methods would not be complete without generating a single multivariate VECM that incorporates all potentially integrated markets. Several versions of a multivariate model have been fitted for this purpose and are discussed next.

Based on the results of the bivariate analysis it can be seen that series 1 and 2 are clearly not cointegrated with the original series. Both series are very distinct from the others and would realistically never be placed in a multivariate framework with them. Moreover, the Johansen trace test estimator selects 0 as the number of cointegrating equations in this multivariate framework containing all twelve series. This strongly conflicts with our expectations, the outcome of the maximum eigenvalue statistic (rank 5) and what is suggested by the information criteria (rank 2 or rank 5). Hence it is more sensible to drop these two series from the multivariate VECM and conduct the analysis with the original series and series 3 through 11. The optimal lag length for this analysis is found to be 2 and the Johansen trace test indicates 7 cointegrating relationships. The stability condition test finds 3 unit eigenvalues, which is the expected value with $K = 10$ and $r = 7$. The autocorrelation test does suggest some autocorrelation in the residuals, which implies a potential misspecification of the model. However, a more significant concern is that when running this analysis

the difficultly of correctly interpreting the results substantially increases without any predetermined constraints.

By default, the Johansen normalisation restrictions are imposed on the cointegrating relationships. However, this simply means that the first seven variables, the original series and series 3 through 8, are set to 1 in one specific cointegrating relationship and 0 in all other six relationships. This results in an analysis where the original series is only included in the first cointegrating relationship and excluded from all the others. This clearly violates the original specification of chapter 4, where the original series is present in all predefined cointegrating relationships. One could attempt to fix the parameters by imposing custom constraints on the multivariate VECM. However, there is no way to know a priori which series should be included in which equation. Normalizing to the original series still leaves the question of which other series to include in every individual equation. Even in this theoretical example with generated series this becomes a significant obstacle for the correct interpretation of the outcome. It should be noted that with r cointegrating equations the minimum number of restrictions required to correctly determine the remaining parameters is r^2 (Johansen 1995).

Normalising each cointegrating equation with respect to the original series provides 7 constraints, leaving 42 constraints to be determined and resulting in under identification of the betas. It is possible to force reporting of the parameters to be estimated when the parameters in beta are not fully identified. Reporting the full results of this procedure would produce a table spanning several pages, which might mystify rather than inform the reader. Every individual series has estimates for 7 adjustment speed parameters and 10 lagged differences, producing 170 parameter estimates. Additionally, each of the seven cointegrating equations features 9 parameter estimates for a total of 233 parameter estimates in the full model. Suffice it to say that the results are difficult to interpret and rather than report the entire table the result are briefly summarized here.

Regarding the cointegrating equations there are a number of unusual results. Series 10 is included in equations 1, 4, 5, 6 and 7 and it is the only parameter estimate that is significantly different from zero for equations 1, 4 and 5 (implying that all three are essentially describing the same cointegrating relationship). Furthermore, series 5 and 9 are both included in three of the seven cointegrating equations, while series 6 is not included in a single one. Since series 6 is designed to be strongly integrated with the original series, while series 9 is only integrated with it to a very weak degree, this is unusual to say the least. The parameter estimates for the adjustment speeds and lagged differences also reveal a number of surprising results. For both series 5 and 6 all seven adjustment speed parameters are significantly different from zero, implying that their price development is determined by no less than seven long-run relationships. Moreover, none of the seven adjustment speed parameters are significantly different from zero for series 10, which is odd as it is included as a parameter estimate that is significantly different from zero in five of the seven cointegrating equations. The results also suggest that the development of series 10 is fully determined by the lagged differences of series 3. There are several more perplexing outcomes of the analysis, but it should be abundantly clear at this point that this version of the multivariate model does not produce useable results and the discussion will be ended at this point for the sake of brevity.

An alternative method for fitting the multivariate VECM despite the above difficulties is to only include the original series and series 5 through 9. These five generated series have well defined parameters and include a cointegrating relationship with the original series by construction. Furthermore, there are no deviations from this predetermined relationship due to shocks, government intervention or other unforeseen eventualities. When using the default Johansen normalisation the results of the fitted model will again be difficult to interpret and not confirm to the original specification in chapter 4. In order to generate workable results after fitting the VECM one thus has to add custom constraints, which are straightforward to determine in this theoretical case. All five individual generated series are cointegrated with the original series without any influence from the other series, which should result in a total of five cointegrating relationships. The optimal lag length is found to be 2 and the number of cointegration relationships is indeed found to be 5 with the Johansen trace test, implying that the minimum number of constraints is equal to 25. Hence, we can impose the constraint that one of the generated series is 1 in each cointegrating equation and exclude the other four generated series by setting them to 0 in each separate equation. This ensures that in every cointegrating equation only one unique generated series is included along with the original series. The result is 5 constraints per cointegrating equation, for a total of 25 (r^2) constraints. Both specification tests indicate that this model is correctly specified (i.e. correct number of cointegrating equations and no signs of autocorrelation in the residuals). The parameter estimates are reported in table 5.5, which follows the same format as table 5.3 for the bivariate results to enable a direct comparison.

Table 5.5: Parameter estimates for multivariate VECM with custom constraints

Series	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$
5	0.186 (0.066) ***	-	0.014 (0.027)	-	-0.679 (0.064) ***	-1.000 (0.003) ***
6	0.099 (0.074)	-	-0.010 (0.060)	-	-0.321 (0.039) ***	-0.995 (0.007) ***
7	0.157 (0.070) **	-	0.081 (0.066)	-	-0.165 (0.031) ***	-1.006 (0.013) ***
8	-0.037 (0.070)	-	0.213 (0.070) ***	-	-0.100 (0.027) ***	-0.970 (0.027) ***
9	0.038 (0.073)	-	0.327 (0.065) ***	-	-0.062 (0.013) ***	-1.063 (0.052) ***

The parameter estimates for β_6 are, as in the bivariate case, quite accurate. Unfortunately, the estimates for the adjustment speed parameter β_5 are less accurate, e.g. its value for series 9 is twice as large as the original specification. However, the estimates still follow the predetermined pattern of integration strength, decreasing from series 5 to 9. Furthermore, the estimates for β_1 to β_4 are very dissimilar from the original specification, with only β_3 being correctly estimated for series 8 and 9. Since the optimal lag length was determined to be two, β_2 and β_4 are not included in the specification of the VECM. The adjustment speed parameter estimates for the 5 different cointegrating relationships are not found to be significantly different from zero for the original series, which is in line with expectations as explained in the bivariate case. For the other series one should only find adjustment speed parameter estimates that are significantly different from zero for the parameter relating to the cointegrating relationship which includes the series under consideration. This expectation holds for all series except for series 9, which has a significant adjustment speed

parameter estimate for both the cointegrating relationship featuring itself and the one featuring series 8. This should obviously not be the case as series 9 is not included in the cointegrating relationship between series 8 and the original series.

The 95% confidence intervals are again used to determine whether the parameter estimates that were found to be significantly different from zero are similar to the true parameter values. The intervals for series 5 to 9 are given in table A.2 of the annex. The parameter estimates for β_5 in series 5 and series 9 turn out to differ in a statistically significant manner from the true values. This implies that the adjustment speed parameter, which indicates the strength of the cointegrating relationship, is incorrectly estimated for these two series. It is particularly striking that the estimate differs to a statistically significant degree from the original specification for series 5, since this is the series that features the strongest degree of integration. In the bivariate analysis all the parameter estimates that were statistically significantly different from zero for series 5 to 9 were in line with the true values. This apparently no longer holds for the multivariate analysis, increasing the likelihood of drawing incorrect conclusions.

The fitted model is again far from perfect, despite imposing a number of theoretical constraints. These constraints were only possible due to the fact that we are dealing with artificially generated data with known values for the true parameters. Using real world data, with unknown specifications, unknown market interactions and a myriad of potential underlying short and long run relationships, in a multivariate VECM will thus likely be even more difficult to correctly fit and interpret. Overall it can be concluded that the bivariate VECMs provides reasonable results, despite the difficulties associated with ascertaining the correct number of lags and analysing series that feature substantial amounts of government intervention or other deviations. The multivariate analysis does not seem to provide useful information about the underlying relationships; it is difficult to interpret and only provides somewhat useful results after imposing custom restrictions on the model and reducing the scope of the analysis by limiting the number of series to be included. It should be noted that the above results were obtained by directly analysing the price levels. An alternative strategy has also been attempted where the prices were converted to their natural logarithms. However, this did not significantly change any of the preceding conclusions and the fitted multivariate model suffered from the same problems highlighted in the above discussion.

5.2 Cluster analysis

Two different analyses will be conducted, one with only the most basic number of variables and one with an additional number of variables. This will highlight the importance of selecting an appropriate set of variables, since this affects the final outcome of the clustering procedure. The first analysis will only include the four basic moments (mean, variance, skewness and kurtosis) of each series, for a total of four variables. The second analysis will expand this basic set of variables. Moreover, as discussed in chapter 3, the variables will be standardized to prevent variables with a high degree of variance from dominating the final outcome. The most straightforward way of doing this is to subtract the mean of a variable and divide by the standard deviation of the variable. Since the primary interest is in the evolution of the series over time, a first step is to analyse the four moments of each series. The first moment is the arithmetic mean, the second moment is the variance, the third moment measures the skewness of the distribution and the fourth moment is the kurtosis. Table 5.6 displays the variable values for the basic cluster analysis, which were calculated using Microsoft Excel 2010.

Table 5.6: Standardized variables for basic cluster analysis

Series	Mean	Variance	Skew	Kurtosis
1	-0.166	-1.764	-0.640	0.075
2	1.824	-1.677	-2.735	1.646
3	-1.769	-0.487	1.173	0.638
4	-1.871	-0.131	0.333	-0.308
5	0.504	1.242	0.624	1.023
6	0.476	1.025	0.356	-0.017
7	0.304	0.692	0.019	-0.955
8	0.087	-0.026	0.247	-0.893
9	-0.222	-0.156	0.575	-0.041
10	-0.023	-0.242	-0.484	-1.682
11	0.326	0.262	-0.144	-0.759
Original	0.529	1.264	0.676	1.272

Linkage criteria under consideration are single linkage, complete linkage, average linkage, median linkage and Ward's linkage. The distance metrics applied are the L_2 metric (squared for median linkage) and the Canberra metric. Based on the discussion in chapter 3 complete linkage might be expected to perform best, as it agglomerates time series in clusters by adding time series that are most similar to the most dissimilar time series already in the cluster. This ensures that the time series that is added to the cluster is relatively similar to all other time series already present in the cluster. However, average/median linkage can potentially outperform this procedure due to the fact that it takes all time series within clusters into account when agglomerating clusters and not just the most extreme time series within a cluster (as is the case with single and complete linkage). Based on the data generation procedure one would expect to find a high degree of similarity between the original series and series 5, as this series was designed to exhibit the strongest degree of integration with the original series. Hence, these two series should form a cluster very rapidly.

A visual review of the dendrograms of the five linkage criteria revealed that average linkage combined with the L_2 metric provides the most useful information for this basic cluster analysis. The dendrogram for average linkage with the L_2 distance metric is given in figure 5.1. The Calinski and Harabasz index and the Duda and Hart index stopping rules are somewhat inconclusive, but both suggest an optimal of five clusters. Based on the dendrogram five distinct clusters seems likely, as after this the distance before the next cluster is joined increases substantially. The red line in figure 5.1 indicates the optimal number of clusters. The overall structure of the dendrogram is reasonably in line with a priori expectations. Both the unrelated markets and the fixed regimes (series 1 to 4) have been separated from the other time series. Table A.3 of the annex provides per cluster averages of the four unstandardized moments. Despite not providing data that can be compared statistically, the table does highlight the differences between the clusters. This very basic analysis thus already provides usable information. It should be noted that the results of the median linkage criterion are identical in terms of the clusters that are formed. Additional variables will be included to refine the analysis and enable an even clearer distinction between related and unrelated series.

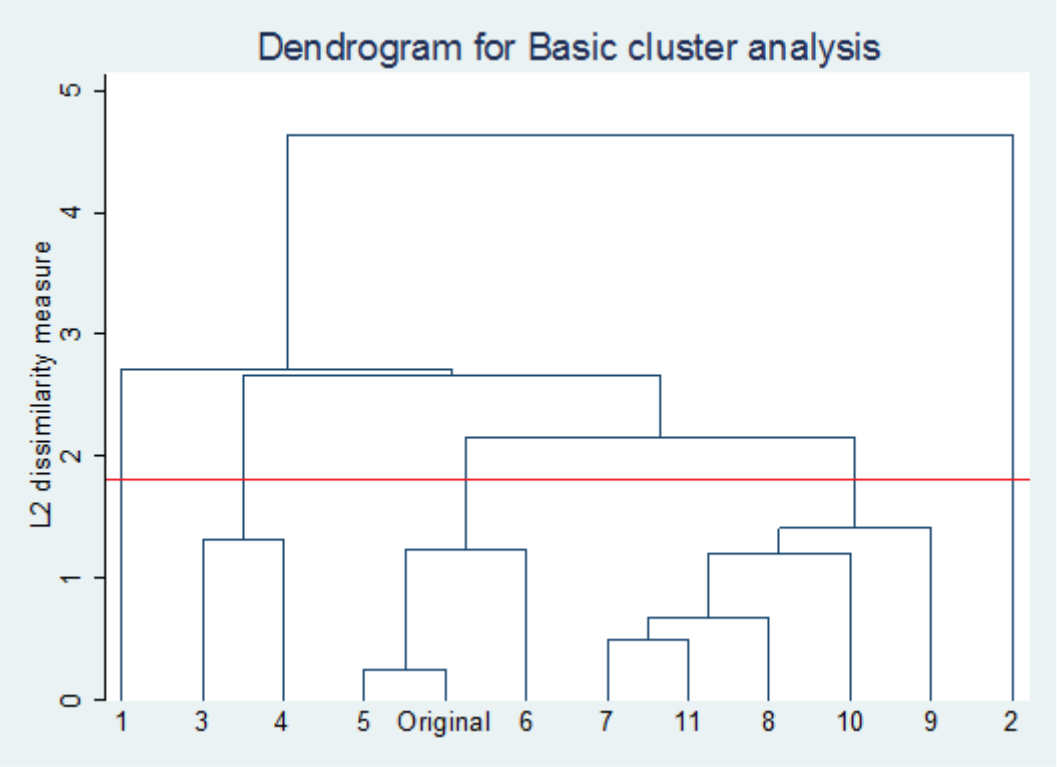


Figure 5.1: Dendrogram of the basic cluster analysis (Average linkage - L_2)

Determining which variables have to be added to the analysis to achieve the sought after refinement mentioned above is a subjective process and draws on the researcher's judgement. What follows is a description of the additional variables that were included in the analysis for this thesis (see table 5.7 for an overview of all variables). It is important to keep in mind that this discussion is not exhaustive and other, perhaps more appropriate variables, might further improve the analysis. As with the basic cluster analysis, the variables are standardized. The first addition is the coefficient of variation, which is the ratio of the standard deviation to the mean and it measures the relative volatility of the series. Next, a dummy is added for government intervention. Based on visual inspection of the price development and official government data it should not be difficult to uncover signs of government intervention in the market. Since the nature of the intervention is similar in the generated data (i.e. consumer protection), the series featuring an intervention are grouped in a single cluster. Should there be distinct types of government intervention several different dummies could be used. For the current analysis this implies the creation of a variable where series 3, 4, 10 and 11 receive a value of one, while the other series are given a value of zero. Furthermore, the median and the interquartile range are added as variables. The median is similar to the arithmetic mean, but less sensitive to extreme observations, and the interquartile range, defined as the third quartile minus the first quartile, is a good measure of the range of the series without the influence of extremes.

Additionally, a number of variables have been included that describe the price development of the individual series. The first straightforward measure that can give an indication of the long run development of a time series is the ratio between the last and first observations in the period under consideration. To describe the short run development of prices three additional measures are used. The number of positive price changes relative to the total number of observations, the number of

absolute changes that are larger than 5% relative to the total number of observations and, lastly, the median of the absolute changes from period to period.

Another fundamental property of the different series is the location of the minimum and maximum values. Moreover, the temporal distance between the maximum and the minimum can also provide information on the speed of change and the order of extremes. This provides the final three variables for the more sophisticated cluster analysis. For the first variable the minimum of the series is divided by the temporal location of the minimum. The second variable is calculated in the same manner, using the maximum value of the series instead of the minimum. Finally, the temporal location of the minimum is deducted from the temporal location of the maximum. For series such as the fixed regimes the minimum and maximum occur in many different periods, due to the fact that price levels are fixed for extended periods of time. For situations such as these the first period in which the minimum or maximum occurs is used in the calculations.

Table 5.7: Standardized variables used in advanced cluster analysis

Category	Indicator
Moments	Arithmetic mean
	Variance
	Skewness
	Kurtosis
Other	Coefficient of variation
	Government dummy
	Median
	Interquartile range
Price development	Ratio last/first observations
	Number of price changes >0 relative to total number of observations
	Number of absolute price changes > 5% relative to total number of observations
	Median of absolute changes from period to period
Temporal development	Value of the minimum divided by temporal location of the minimum
	Value of the maximum divided by temporal location of the maximum
	Temporal distance maximum-minimum

Running the cluster analysis with these fifteen variables should bring more clarity to the analysis, as more information is extracted from the price series than with the basic analysis. Once again average linkage combined with the L_2 metric provides the most sensible outcome and the dendrogram is given in figure 5.2. As before, median linkage also performs very well and is particularly good in differentiating between series 1-4 and the others. However, it is not as capable as average linkage when it comes to finding close relationships with the original series, as it creates a cluster consisting exclusively of the original series and series 5. Therefore, average linkage has been selected as the preferred linkage criterion in the analysis.

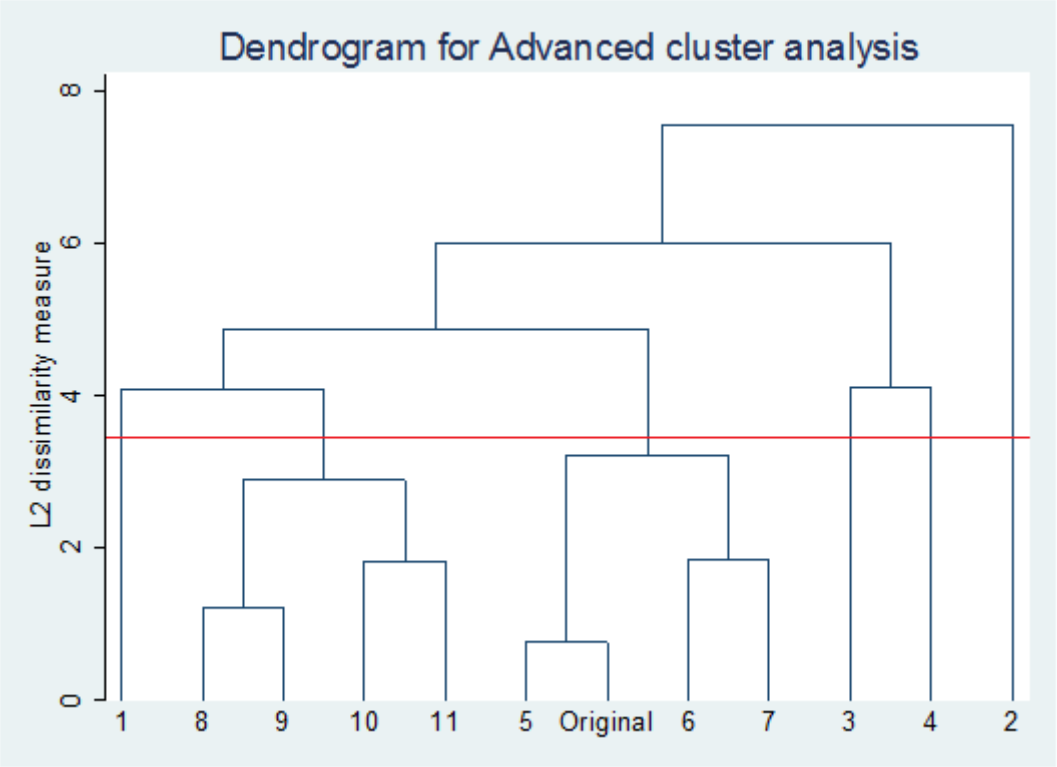


Figure 5.2: Dendrogram of advanced cluster analysis (Average linkage - L_2)

The stopping rules are again inconclusive, but the Duda and Hart index in particular strongly suggests that five clusters are most appropriate for this analysis. However, the dendrogram reveals that this is a rather unnatural point to stop the clustering procedure, as the distance between the five and six cluster solutions is miniscule. Hence, in this case the appropriate number of clusters is difficult to determine and once again the researcher's judgement comes into play. Based on the a priori knowledge about the time series it should be clear that the unrelated markets (series 1 and 2) are distinct from the others. Utilising this information in combination with the dendrogram leads to the conclusion that the six cluster solution provides the most accurate description of the data, as indicated in figure 5.2 by the red line. The outcome of the analysis reveals that series 5, 6 and 7 are similar enough to the original series to be placed within the same cluster. This is in line with expectations as these series are integrated to a very strong, strong and medium degree respectively. Series 8 and 9 only showed a very small degree of integration with the original series and are placed in the same cluster as series 10 and 11, which featured medium integration with some sort of deviation that caused the integrating relationship to not hold at all times. Series 1, 2, 3 and 4 are all placed in individual clusters and are as such, based on the variables used in the analysis, considered to be different from the other series. This is again in line with expectations as series 1 and 2 represent two unrelated markets and series 3 and 4 are the fixed regimes. The difference with the basic cluster analysis is that series 3 and 4 are now placed in separate clusters, increasing the total number of clusters from five to six. Additionally, series 7 changes cluster and joins the cluster containing the original series. Table A.4 of the annex provides the per cluster averages of the four unstandardized moments. The table only includes the four moments to allow comparison with the outcomes of the basic cluster analysis in table A.3. Please keep in mind that these numbers are not comparable in a statistical manner and only serve to illustrate differences. The accuracy of the final

outcome has improved somewhat, but there are no major changes, since the basic cluster analysis was already in line with the a priori expectations to a reasonable degree.

As discussed in section 5.1 a common approach while conducting cointegration analysis is to convert the price levels to their natural logarithms. To see how this impacts the outcome of the cluster analysis, the above analysis is repeated with all fifteen variables recalculated using the natural logarithms of the prices. To ensure the results are directly comparable the analysis is conducted using the same linkage criterion and distance metric as before, namely average linkage combined with the Euclidian distance. The outcome is reported in figure 5.3 and is exactly the same as the previous analysis in terms of cluster outcome. The optimal number of clusters is again six and all series remain agglomerated within the same groups.

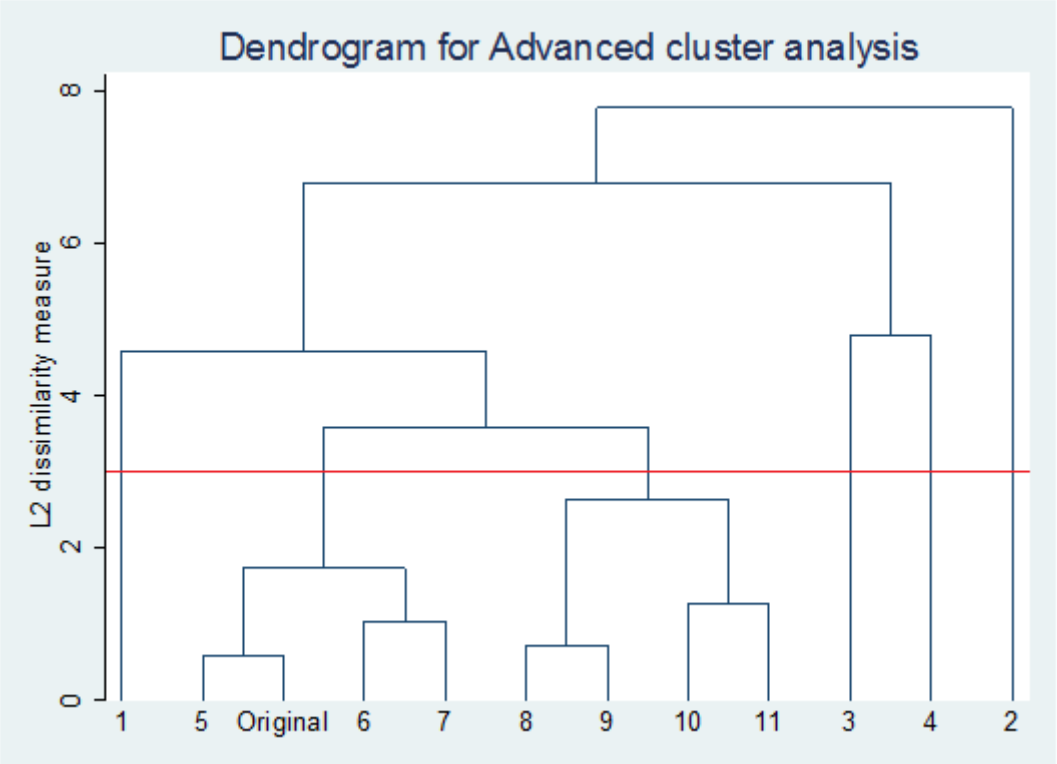


Figure 5.3: Dendrogram of advanced cluster analysis using natural logarithms (Average linkage - L_2)

Compared to the previous analysis, series 1 seems to be more distinct from the original series and series 5 through 11, which is a desirable outcome. However, since the final cluster solution is the same, there is no clear way of determining which analysis is more appropriate. It appears that it does not matter greatly whether one chooses to use prices in levels or natural logarithms. This indicates that the analysis is both flexible and robust to changes in scale.

5.3 Sensitivity analysis

One of the primary objectives, which underline the potential advantage of cluster analysis, is assessing the method's robustness with regard to missing data. As mentioned numerous times in the preceding chapters, cointegration analysis requires a complete dataset and cannot be executed should there be missing observations. Cluster analysis does not suffer from this very stringent requirement; as long as the variables used in the cluster analysis itself are complete, there is no problem. Introducing missing observations into the dataset does not impact the ability to calculate the fifteen variables specified in the preceding section. It merely implies that the values of the variables will change to a certain degree. To investigate the effect of missing data on the final outcome of the clustering solution, three simulations will be conducted in this section. All data points of the generated series receive a chance of being converted to a missing observation, with probabilities of 5%, 10% and 20% respectively. The original series will not feature any missing observations, as this kind of world price data is typically complete and straightforward to acquire. Table 5.8 provides an overview of the number of missing data points per series, with the percentages in parentheses. The analysis will be performed using the average linkage criterion and the L_2 distance metric as before.

Table 5.8: Simulated missing observations per series

Series	1	2	3	4	5	6	7	8	9	10	11
Missing (5%)	10 (5)	7 (3.5)	12 (5.9)	12 (5.9)	10 (5)	10 (5)	15 (7.4)	11 (5.5)	14 (6.9)	10 (5)	13 (6.4)
Missing (10%)	24 (11.9)	20 (9.9)	20 (9.9)	17 (8.4)	21 (10.4)	16 (7.9)	20 (9.9)	18 (8.9)	14 (6.9)	14 (6.9)	23 (11.4)
Missing (20%)	32 (15.8)	43 (21.3)	36 (17.8)	47 (23.3)	34 (16.8)	47 (23.3)	34 (16.8)	39 (19.3)	45 (22.3)	28 (13.9)	41 (20.3)

The results of the 5% and 10% missing data points cluster analyses are exactly the same as before and will therefore not be discussed any further. The 20% missing observations simulation does feature a slightly different outcome, as can be seen in figure 5.4. The stopping rules combined with a visual inspection of the dendrogram lead to the conclusion that the six cluster solution is once again most appropriate. Series 7 changes cluster in this simulation and leaves the group containing the original series. The analysis is still capable of identifying series 5 and 6 as being very closely related to the original series. Hence, some detail is lost in the analysis but the results are still very much in line with the underlying structure of the generated series. Based on this simple simulation it can be concluded that cluster analysis is indeed quite robust against missing observations. Even when the dataset under analysis features 20% missing observations the results are virtually unchanged.

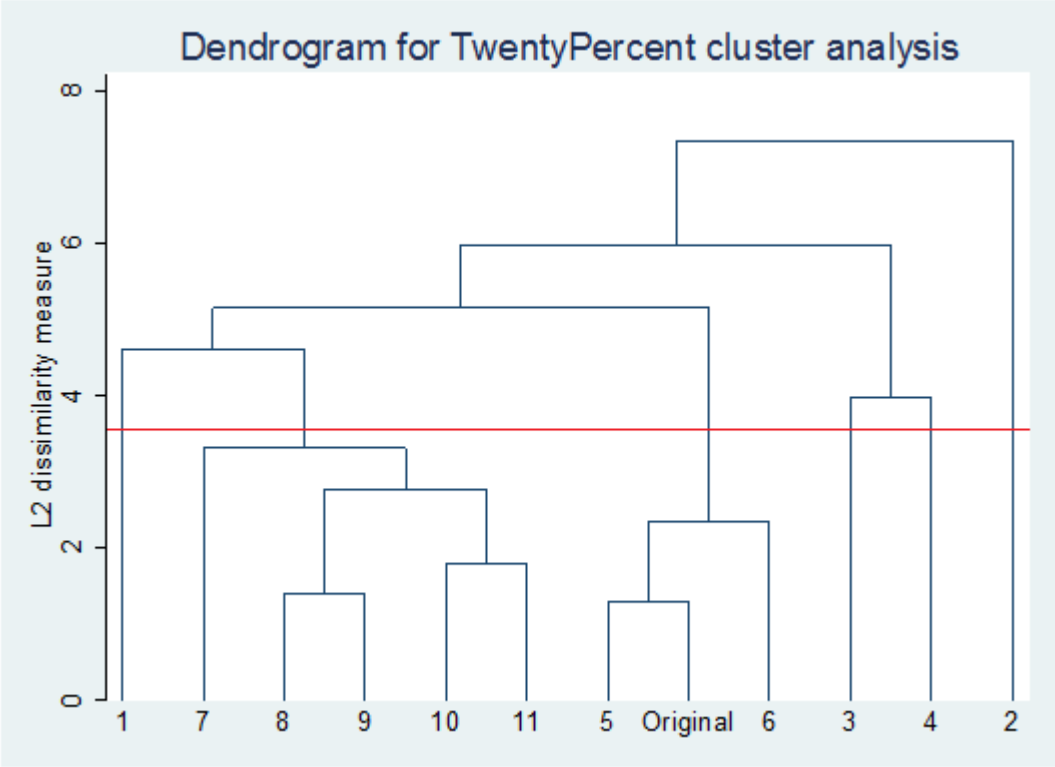


Figure 5.4: Dendrogram of advanced cluster analysis with 20% missing observations (Average linkage - L_2)

6. Conclusions and recommendations

This final chapter brings together the information provided in the preceding chapters. The first section provides the main conclusions and answers the research questions. The second section covers the discussion and provides a critical reflection on the results obtained during this study. Lastly, the third section offers suggestions for further research.

6.1 Conclusion

The literature review conducted for the theoretical framework revealed that market integration is a complex concept. The most basic version of market integration is known in the literature as the Law of One Price, which is dichotomous in nature. Markets are either fully integrated or not at all. Measuring the degree of integration on a continuous scale is a more sensible approach, which is achieved by using the price transmission ratio. It is practical to separate price transmission into three components measuring the degree of co-movement, the speed of adjustment and the asymmetry in price response respectively. Furthermore, one needs to keep the distinction in mind between competitive equilibrium and market integration. This directly relates to the available data, as it requires both price data and information on trade volumes. In this study the focus was exclusively on price data and it is thus not possible to make the distinction. There are also numerous factors that can inhibit market integration, such as transaction costs, market power and trade policy. To analyse the degree of market integration, economists over the years moved from bivariate correlation coefficients to cointegration analysis. Cointegration has become the primary tool used to study market integration in practice. The method enables the researcher to find estimates for both short and long term price adjustment, while also indicating the speed of adjustment. However, the method assumes stationarity of transactions costs and cannot handle datasets which feature missing observations or other data related problems.

Applying the cointegration methodology to the generated data provided a good benchmark against which to test cluster analysis. In the bivariate framework all generated series were separately analysed in conjunction with the original series. Cointegrating relationships were not found when analysing the two unrelated markets, confirming to expectations. For the remaining nine series the bivariate analysis revealed cointegrating relationships, indicating the presence of a long-term relationship between each of the individual series and the original series. The original specification featured a total of three lagged terms, which was only correctly detected for the very strongly integrated series (series 5) and the series featuring medium integration and a price shock (series 11). The direct result of this was that the short term price adjustments, as captured by the lagged differences in the error correction model, were not fully in line with a priori expectations. Furthermore, the adjustment speed parameter was found to be significantly different from the original specification for both series that were integrated to a medium degree with the original series and featured some form of deviation (series 10 and 11). The implication being that when markets deviate from their ideal relationships by unexpected events such as price shocks or government interventions, bivariate cointegration analysis struggles to find the true parameter values. It was also attempted to place all series into a combined multivariate framework that would enable the analysis of the market as a whole. Unfortunately, the fitted model featuring 10 series (the unrelated markets were dropped from the analysis) proved to be difficult to interpret and irreconcilable with the original specification of the generated data. By restricting the number of series to six and imposing a

strict set of constraints, somewhat useable results were obtained. However, it was only possible to apply the constraints due to the fact that the series followed a predetermined and known relationship. When analysing real world data one cannot possibly know the true underlying structure of the data or the potential myriad of relationships between different markets, making the analysis and correct interpretation of the model truly daunting. Moreover, the results that were obtained with the restricted multivariate model were unable to reveal the correct short term dynamics and overestimated the adjustment speed parameter for series 5 and 9 to a statistically significant degree. Overall it can be concluded that cointegration analysis in a bivariate setting works rather well, provided the series do not feature deviations from their ideal relationships. Analysing the different time series in this particular manner does obfuscate any potential cointegrating relationships that are more complex and feature multiple series. Attempting to successfully analyse large numbers of markets in a multivariate model simultaneously proves to be elusive, if not impossible.

Next, cluster analysis was investigated as a possible alternative to cointegration analysis, with a specific focus on analysing large numbers of markets simultaneously and the potential robustness of the method to flawed data. The analysis of the data revealed that cluster analysis is indeed very capable of differentiating between time series that develop in a similar manner and those that do not. The advanced cluster analysis, which utilized fifteen variables in an attempt to extract that maximum amount of information from the time series, was able to identify the very strongly integrated series, the strongly integrated series and the medium integrated series as closely resembling the original series. Moreover, the analysis placed each of the two unrelated markets and two fixed regimes into four separate clusters. The two series featuring a weak degree of integration (series 8 and 9) and the two series featuring deviations from the long-term relationship (series 10 and 11) were jointly placed in a separate cluster. The series thus did not resemble the original series to a close enough extent to be considered integrated with it. The bivariate cointegration analysis was able to detect market integration in this case. Considering the weakness of the relationship this result is not unsurprising and it seems that cluster analysis is not able to detect very weak degrees of market integration. Measuring prices either in levels or natural logarithms was found to not influence the analysis, resulting in exactly the same clustering outcome. Finally, three sensitivity analyses were conducted to investigate the robustness of cluster analysis to missing data. The first featured 5% missing data, the second 10% and the third 20%. The analyses revealed that the results of the final clustering outcome were only affected in the case of 20% missing observations. The series that featured a medium level of integration (series 7) was no longer clustered with the original series. Hence, the accuracy of the procedure was affected to a small degree. When considering the substantial amount of missing data in each individual time series in this simulation, the results can still be considered rather accurate. The implication is that cluster analysis is a particularly useful tool for economists when working with empirical data that is less than perfect. It is thus a powerful and practical tool that is capable of revealing relationships between time series in a straightforward manner and is especially suitable for analysing large numbers of market simultaneously.

Overall it can be argued that cluster analysis can indeed be an appropriate tool for analysing market integration. Whether or not the method should be applied depends on the objectives of the researcher. If a limited number of markets are to be investigated and the data features no major shortcomings, the cointegration approach would be the preferred method. It offers information on short and long term effects and also enables the researcher to investigate the speed of adjustment, none of which are possible with cluster analysis. However, when dealing with flawed data the

accuracy of the parameter estimates provided by the cointegration method could give a false sense of exactness and, additionally, could result in erroneous conclusions about the true nature of the underlying market relationships. Furthermore, when analysing a larger number of time series simultaneously cointegration analysis fails to provide workable outcomes. Cluster analysis thus provides an alternative in empirical cases with flawed data or larger numbers of markets and enables researchers to draw conclusions in situations where cointegration analysis would be unsuccessful or provide nonsensical outcomes.

6.2 Discussion

The most obvious shortcoming of this thesis is that all analyses were conducted on artificially generated data. The data were generated based on empirical time series data of global wheat prices, but the overall structure was kept as simple as possible. By not introducing overly complex lag structures and cointegrating relationships the analyses were straightforward to do and interpret. Since the goal of the thesis is mainly to demonstrate the usefulness of cluster analysis in the analysis of market integration this was an acceptable strategy. Moreover, by utilizing the generated data it was possible to directly compare the outcome of the cointegration method to the predefined parameter values. This revealed that cointegration analysis struggles to find the true parameter values in a multivariate setting and in the case of interventions. Hence, the use of artificial data also offered a unique advantage that was crucial for the conclusions of this thesis. However, for cluster analysis to become a functional tool to economists in the analysis of market integration it should undergo rigorous testing with real world data.

Another shortcoming relates directly to the method itself. As mentioned numerous times throughout this thesis, cluster analysis is an inherently subjective method that does not allow one to draw any statistically significant conclusions. This also makes it difficult to compare results between researchers. It is thus of utmost importance that the method is developed into a standard toolset that is applied in a consistent manner. This includes developing a fixed basic set of variables and deciding which linkage criterion and distance metric to utilize. The results in this thesis indicate that average and median linkage combined with the l_2 metric offer the best outcomes. It is advisable to use the dendrograms of both linkage criteria and compare them to what can reasonably be expected. Since the method does not offer any statistical tests it is important to combine cluster analysis with additional information and theory. Simply providing the outcome of the cluster analysis will never provide sufficient evidence for any pattern of market integration. It is important that researchers combine the method's outcomes with practical and theoretical information on trade patterns, government policy, data on price shocks etc. Cluster analysis is thus in itself not sufficient, but merely provides supportive evidence for making well-reasoned arguments about market integration.

Furthermore, it should be noted that cluster analysis does not allow the researcher to make any distinction between the degree of co-movement in the short and long term or draw any conclusions about adjustment speeds. The outcome of the analysis is binary, markets are integrated or they are not. However, when considering subpar empirical datasets and the difficulties associated with correctly interpreting a multivariate VECM, it does offer a workable alternative for research on market integration. Without cluster analysis it might, in some cases, not even be possible to conduct any form of analysis with regard to market integration at all. Even when it is possible to fit a multivariate VECM by data interpolation and other cleaning methods, the parameter estimates obtained in this manner are unlikely to represent the true values and may lead to erroneous

conclusions. In a way cluster analysis takes a step back and does not pretend to have all the answers. Rather than offer a false sense of precision with exact parameter estimates, it provides clarity on market integration in the simplest of manners. The rest is up to the researcher.

6.3 Further research

As discussed in the previous sections cluster analysis is not meant to replace existing methods. It merely offers a tool to investigate large numbers of markets simultaneously, which might also feature shoddy data. However, since cluster analysis is a completely novel approach in the analysis of market integration many avenues for further research are wide open. One of the fundamental objectives of further research should be to develop a standard set of variables, which can extract the maximum amount of information from the time series under consideration. The fifteen variables used in this thesis can prove to be a suitable starting point, but more research is required to determine an optimal and consistent basic set of variables. It is likely that a more sophisticated set of variables will be able to further refine the method's accuracy, as was demonstrated by the difference between the basic and advanced cluster analyses. It bears repeating that a different set of variables can, and most probably will, lead to different outcomes of the final clustering solution. For the development of a standard set of variables it is important to keep in mind that the method's relatively uncomplicated nature is one of its strengths and the set of variables to be used in the analysis should therefore not become too large or feature overly complex variables.

Another major issue is the lack of any definitive stopping rule. This adds further subjectivity to the method and can potentially lead different researchers to draw contrasting conclusions based on the same time series data. A combination of multiple existing stopping rules can be applied, as was done in this thesis, but perhaps it is more sensible to develop a new stopping rule or decide on a single rule to determine the optimal clustering outcome. It is likely that no matter which stopping rule is adopted as the standard approach, the researcher's judgement will keep playing a substantial role in determining the final outcome. This is one of the reasons why it was argued in the preceding sections that cluster analysis in itself is never enough, it requires additional evidence to make a decisive judgement about the nature of market integration.

References

Abdel-Latif, AM & Nugent, JB 1996, 'Transaction cost impairments to international trade: Lessons from Egypt', *Contemporary Economic Policy*, 14(2), 1-14.

Bakucs, LZ, Brümmer, B, Cramon-Taubadel, S von & Ferto, I 2012, 'Wheat market integration between Hungary and Germany', *Applied Economics Letters*, 19(8), 785-788.

Barrett, CB & Li, JR 2002, 'Distinguishing Between Equilibrium and Integration in Spatial Price Analysis', *American Journal of Agricultural Economics*, 84(2), 292-307.

Barrett, CB 2001, 'Measuring Integration and Efficiency in International Agricultural Markets', *Review of Agricultural Economics*, 23(1), 19-32.

Bellemare, MF, Barrett, CB & Just, DR 2013, 'The Welfare Impacts of Commodity Price Volatility: Evidence from Rural Ethiopia', *American Journal of Agricultural Economics*, 95(4), 877-899.

Cudjoe, G, Breisinger, C & Diao, X 2010, 'Local impacts of a global crisis: Food price transmission, consumer welfare and poverty in Ghana', *Food Policy*, 35(4), 294-302.

Datta, DD & Du, W 2012, *Nonparametric HAC Estimation for Time Series Data with Missing Observations*, International Finance Discussion Paper Number 1060, Board of Governors of the Federal Reserve System, viewed 6 December 2014, http://www.federalreserve.gov/pubs/ifdp/2012/1060/ifdp1060.pdf.

Dougherty, C 2011, '13. Introduction to Nonstationary Time Series', *Introduction to Econometrics*, 4th edition, Oxford University Press, 463-513.

Fackler, PL & Goodwin, BK 2001, 'Spatial price analysis', in BL Gardner & GC Rausser (eds.), *Handbook of Agricultural Economics*, Elsevier Science, 971-1024.

Fackler, PL & Tastan, H 2008, 'Estimating the Degree of Market Integration', *American Journal of Agricultural Economics*, 90(1), 69-85.

FAO 2014, *FAOSTAT*, viewed 4 November 2014, http://faostat.fao.org/site/342/default.aspx.

FAO 2015, *Food Price Monitoring and Analysis Tool*, viewed 20 January 2015, http://www.fao.org/giews/pricetool/.

Ghosh, M 2003, 'Spatial Integration of Wheat Markets in India: Evidence from Cointegration Tests', *Oxford Development Studies*, 31(2), 159-171.

Gordon, AD 1999, *Classification*, 2nd edition, Chapman & Hall/CRC.

Gouel, C 2013, 'Optimal food price stabilisation policy', *European Economic Review*, 57, 118-134.

Goychuk, K & Meyers, WH 2014, 'Black Sea and World Wheat Market Price Integration Analysis', *Canadian Journal of Agricultural Economics*, 62(2), 245-261.

Granger, CWJ 2004, 'Time Series Analysis, Cointegration, and Applications', *The American Economic Review*, 94(3), 421-425.

Hair, JF, Anderson, RE, Tatham, RL & Black, WC 1998, 'Chapter 9 Cluster Analysis', *Multivariate Data Analysis*, 5th edition, Prentice Hall International Inc, 469-518.

Hobijn, B, Franses, PH & Ooms, M 1998, *Generalizations of the KPSS-test for Stationarity*, Report no. 9802/A, Econometric Institute, Erasmus University Rotterdam.

Johansen, S 1995, *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press.

Krugman, PR, Obstfeld, M & Melitz, MJ 2012, 'Chapter 10 – The Political Economy of Trade Policy', *International Economics: Theory & Policy*, 9th edition, Pearson Education Limited, 249-285.

Lattin, JM, Carroll, JD & Green PE 2003, '8 - Cluster Analysis', *Analyzing Multivariate Data*, 1st edition, Brooks/Cole, 264-310.

Löfgren, H & El-Said, M 2001, 'Food subsidies in Egypt: reform options, distribution and welfare', *Food Policy*, 26(1), 65-83.

Meyer, J & Cramon-Taubadel, S von 2004, 'Asymmetric Price Transmission: A Survey', *Journal of Agricultural Economics*, 55(3), 581-611.

Milligan, GW & Cooper, MC 1985, 'An Examination of Procedures for Determining the Number of Clusters in a Data Set', *Psychometrika*, 50(2), 159-179.

Minot, N 2011, *Transmission of World Food Price Changes to Markets in Sub-Saharan Africa*, Discussion Paper 01059, International Food Policy Research Institute (IFPRI), viewed 24 November 2014, http://www.ifpri.org/sites/default/files/publications/ifpridp01059.pdf.

Mukhtar, T & Ishfaq, M 2009, 'WTO and Pakistan's Agriculture: a Price Integration and Welfare Analysis for Wheat', *Journal of Economic Cooperation and Development*, 30(1), 59-86.

Nielsen, B 2001, 'Order determination in general vector autoregressions', *IMS Lecture Notes–Monograph Series*, 52, 93-112.

Perkins, DH, Radelet, S & Lindauer, DL 2006, 'Chapter 16 - Agriculture', *Economics of Development*, 6th edition, W.W. Norton & Company, 607-650.

Rapsomanikis, G, Hallam, D & Conforti, P 2006, 'Market integration and price transmission in selected food and cash crop markets of developing countries: review and applications', in A Sarris & D Hallam (eds.), *Agricultural commodity markets and trade. New approaches to analyzing market structure and instability*, Commodities and Trade Division of FAO, Rome, 187-217.

Sharma, R 2003, 'The Transmission of World Price Signals: the Concept, Issues and some Evidence from Asian Cereal Markets', in *Agricultural Trade and Poverty – Making Policy Analysis Count*, Organisation for Economic Co-operation and Development, 141-160.

Timmer, CP 1989, 'Food price policy: The rationale for government intervention', *Food Policy*, 14(1), 17-27.

Varela, G, Aldaz-Carroll, E & Iacovone, L 2012, *Determinants of Market Integration and Price Transmission in Indonesia*, WPS6098, The World Bank, viewed 4 December 2014, http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2012/06/19/000158349_20120619102511/Rendered/PDF/WPS6098.pdf.

Verbeek, M 2012, '9 - Multivariate Time Series Models', *A Guide to Modern Econometrics*, 4th edition, John Wiley & Sons Ltd, 338-371.

Vercammen, J 2011, *Agricultural Marketing: Structural models for price analysis*, 1st edition, Routledge.

Volkskrant 2014, 'Russische boycot: prijsdalingen, peerselfies, chaos bij de grens', *Volkskrant*, viewed 23 November 2014, http://www.volkskrant.nl/economie/russische-boycot-prijsdalingen-peerselfies-chaos-bij-de-grens~a3715501/.

World Bank 2015, *GEM Commodities*, viewed 20 January 2015, http://data.worldbank.org/data-catalog/commodity-price-data.

Table A.1: Bivariate analysis - 95% Confidence intervals for parameter estimates

Series	$\widehat{\beta_1}$		$\widehat{\beta_2}$		$\widehat{\beta_3}$		$\widehat{\beta_4}$		$\widehat{\beta_5}$		$\widehat{\beta_6}$	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
5	0.102	0.366	-0.029	0.190	-0.180	0.056	-0.042	0.019	-0.703	-0.441	-1.007	-0.996
6	0.128	0.215	-	-	-0.060	0.076	-	-	-0.319	-0.236	-1.016	-0.990
7	0.074	0.125	-	-	-0.024	0.173	-	-	-0.153	-0.109	-1.028	-0.975
8	0.009	0.050	-	-	0.064	0.313	-	-	-0.085	-0.056	-1.001	-0.899
9	-0.008	0.030	-	-	0.190	0.443	-	-	-0.037	-0.021	-1.210	-0.941
10	0.081	0.184	-	-	-0.020	0.218	-	-	-0.121	-0.050	-0.946	-0.770
11	0.078	0.146	-0.056	0.021	0.441	0.708	-0.290	-0.053	-0.084	-0.037	-1.052	-0.860

Table A.2: Multivariate analysis - 95% Confidence intervals for parameter estimates

Series	$\widehat{\beta_1}$		$\widehat{\beta_2}$		$\widehat{\beta_3}$		$\widehat{\beta_4}$		$\widehat{\beta_5}$		$\widehat{\beta_6}$	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
5	0.056	0.316	-	-	-0.038	0.066	-	-	-0.805	-0.554	-1.007	-0.994
6	-0.045	0.244	-	-	-0.127	0.108	-	-	-0.399	-0.244	-1.009	-0.981
7	0.020	0.295	-	-	-0.049	0.211	-	-	-0.225	-0.105	-1.032	-0.980
8	-0.175	0.101	-	-	0.075	0.350	-	-	-0.154	-0.047	-1.024	-0.917
9	-0.105	0.180	-	-	0.199	0.455	-	-	-0.088	-0.036	-1.166	-0.961

Table A.3: Basic cluster analysis – per cluster average unstandardized moments

Series in cluster	Mean	Variance	Skewness	Kurtosis
1	189.55	1702.05	0.19	-1.16
2	239.82	1851.88	-0.45	-0.78
3, 4	147.77	4212.75	0.62	-1.14
Original, 5, 6	206.45	6776.65	0.56	-1.00
7, 8, 9, 10, 11	196.12	4929.20	0.40	-1.39

Table A.4: Advanced cluster analysis – per cluster average unstandardized moments

Series in cluster	Mean	Variance	Skewness	Kurtosis
1	189.55	1702.05	0.19	-1.16
2	239.82	1851.88	-0.45	-0.78
3	149.06	3906.08	0.75	-1.03
4	146.49	4519.43	0.49	-1.25
Original, 5, 6, 7	205.19	6567.49	0.52	-1.10
8, 9, 10, 11	194.80	4676.49	0.41	-1.38