Time Series Clustering in High Dimensional Cointegration Analysis: The Case of African Swine Fever in China

Rundong Peng, Michigan State University, pengrund@msu.edu.
Mindy Mallory, Purdue University, mlmallory@purdue.edu.
Meilin Ma, Purdue University, mameilin@purdue.edu.
H. Holly Wang, Michigan State University, wanghong@msu.edu.

Selected Poster prepared for presentation at the 2025 AAEA & WAEA Joint Annual Meeting in Denver, CO: July 27-29, 2025

Time Series Clustering in High Dimensional Cointegration Analysis: The Case of African Swine Fever in China

Abstract

Time series data have been extensively utilized in agricultural price analysis, with the Vector Auto-Regressive (VAR) and Vector Error Correction Model (VECM) being foundational tools. Over the past three decades, the availability of disaggregated agricultural commodity price data has increased, resulting in high-dimensional datasets. The efficacy of VECM and Johansen's maximum likelihood test diminishes with increased dimensionality due to exponential growth in the required time series length, implying difficulty in extracting cointegrating relationships in high-dimensional data. This article addresses this challenge by employing time series clustering to reduce data dimensionality. Clusters are formed based on price similarity, dynamically adjusted for specified time period using hierarchical clustering with dynamic time warping. With clustered time series, we extract the mean price of each cluster and apply Johansen's framework to estimate cointegration relationships. Applied to the Chinese hog market before and after the 2018 African Swine Fever outbreak, we show that the cointegrating relationship has changed suggesting less inter-provincial trade. The study identifies clusters based on price similarity and shows the advantages of this method compared to traditional geographical clustering.

JEL Codes: Q02, Q11, Q18

Key words: cointegration, high-dimensional time series models, machine learning, market equilibrium

1 Introduction

Time series data have been widely used in agricultural price analysis. The Vector Auto-Regressive (VAR) and the Vector Error Correction Model (VECM) are workhorses in modeling economic time series. When the linear combination of different time series is stationary, these series are cointegrated. The VECM, which uncovers both short-term and long-term dynamics, is preferred over the VAR, when the series are cointegrated.

The amount of data available for analysis has been growing dramatically since more than 30 years ago when Engle and Granger Engle and Granger (1987) proposed the error correction model on cointegrating series and when Johansen introduced the likelihood ratio test on cointegration (Johansen, 1988). For agricultural commodity prices, the time series data are now available at a more disaggregated level and for more commodities than ever before. The unprecedented number of time series available can form a high-dimensional dataset.

With high-dimensional data, the efficacy of the VECM and the Johanson maximum likelihood test for cointegration is limited. In Johansen's framework, as the number of time series in the VECM grows, the requirement on the length of time series grows exponentially to make sure that the likelihood ratio test holds its asymptotic property. One way to estimate high-dimensional VECM is using penalized least squares (Basu and Michailidis, 2015). This method overcomes the strict requirement of the length of the series. However, due to the nature of penalized least squares, it cannot identify the long-term cointegration relationship directly and fails to generate the confidence interval for coefficients.

One alternative way to address the challenge of high-dimensional data is to reduce the dimension by grouping the large number of series into a few *representative series* to fit the Johansen's framework. For example, instead of using a large number of county- or state-level series, many prior studies on agricultural prices used regional-level price data. The number of regions is typically 3 to 5, and the regions are defined according to their geographical boundaries by the government for administrative purposes (Wang, Chavas, and Li, 2023).

Aggregating data to the region level can reduce data dimension, but does not reflect price patterns in a specific time window. Using time series clustering to specify clusters is a data-driven process on specified periods that completely relies on the patterns of prices.

This article applies time series clustering to reduce the data dimension to address the difficulty of modeling high-dimensional cointegrated time series with VECM. Time series are clustered based on their similarity in a given period. In different periods, a cluster may have different elements. We calculate the centroid (mean of all time series) of each cluster to obtain a new dataset with fewer series. After reducing the dimension to a suitable level, we are able to use Johansen's approach to employ a VECM model that estimates the long-term cointegration relationship directly in the usual way.

We apply a combination of methods in time series clustering to the Chinese hog market over periods before and after the 2018 African Swine Fever (ASF) outbreak. Our time series clustering results reveal distinct separation among the clusters, aligning closely with our understanding of the Chinese hog industry. We analyze the long-term cointegration and short-term price transmission of hog prices among clusters of provinces both before and after ASF. Compared with geographical clustering, our results successfully captured the change in the significance of cointegration relationships due to ASF outbreak.

This article makes contributions to the literature by adopting time series clustering to reduce the dimension of a dataset Among different types of time series clustering methods and parameters that can be used in clustering, we employ a hierarchical clustering approach with dynamic time warping (DTW) which is appropriate for clustering price series because DTW allows for price transmission that may occur over a number of periods.

Our study also provides new information on the price patterns of the Chinese hog market. Provinces with low trade friction tend to have similar price patterns due to competition eliminating spatial arbitrage. We identify clusters based on observed similarity in prices rather than aggregating based on an arbitrary method such as aggregation of provinces to a region. Compared with aggregating the price geographically, our approach is robust to unobserved trade flows and allows us to use fewer clusters with lower within-cluster variation and higher among-cluster variation. This approach sufficiently reduces the dimension of a VECM with Johansen's specification to recover the cointegrating relationships among the clusters. The results of time series clustering are suggestive of groups of provinces with such an absence of spatial arbitrage. Large producers and the government could use this information for resource allocation and policy making.

We conclude from results of the VECM model that ASF had changed the cointegrating relationship among different clusters. Before the ASF, the error correction terms are significant for all clusters suggesting a strong mean-reverting pattern from arbitraging. While after the ASF, we do not observe much significant from error correction terms but one cluster shows high significance in all equations indicating less nation-wide arbitrage. Large producers and the government could use this information for resource allocation and policy making in response to restore the hog market from disease.

2 Related Literature

This article mainly speaks to three strands of literature: studies of agricultural commodity prices, studies of price transmission using VECM, and empirical analysis of agricultural commodities using clustering.

Classical literature studied prices as commodities are transformed over space, time, and form. Bressler Bressler et al. (1970) discusses the factors including transfer costs, processing costs, and storage costs affecting the price, and prices can be viewed as a complex structure reflecting geographic differences, form differences, and time differences. In the US wheat market, transport cost affects spatial price differentials (Roehner, 1996). Transportation cost, infrastructure, and policy have a strong influence on the price of perishable goods (Beghin and Schweizer, 2021). Vachal Vachal (2015) found that the truck contracting for grain shipping in the northern plains of the US is heterogeneous, suggesting that transport costs can be different even within a region. Changes in trade policy and non-tariff measures have caused increases in agricultural commodity prices (Tokgoz et al., 2011). Morrison Paul and MacDonald Morrison Paul and MacDonald (2003) showed that there is a linkage among farm commodity prices, food processing costs, and food prices. Storage patterns of agricultural commodities affect both prices and fluctuations (Mitra and Boussard, 2012).

Research on spatial integration in the supply chain has been summarized by von Cramon-Taubadel and Goodwin von Cramon-Taubadel and Goodwin (2021). Price transmission is studied to link the difference in price and the Law of One Price. The error correction model has been used widely in studying price transmission for its advantage in estimating long-term cointegration relationships. Palaskas Palaskas (1995) showed that even when commodity prices drift apart in the short-run due to producer price policy changes, the market forces causes prices return to their long-run equilibrium.

Classical work (Gardner, 1975) adopted structural and static models in studying price transmission. More recent work used the time series approach and VECM. Espost and Listorti Esposti and Listorti (2013) studied the price transmission across agricultural commodities and space. They showed that temporary trade intervention had limited the impact of price bubbles on price transmission. Hatzenbuehler et al. Hatzenbuehler, Abbott, and Abdoulaye (2017) found that price transmission is directly related to the traceability of the agricultural commodity. Zhou and Dieter (De and Koemle, 2015) used VECM to study price transmission in livestock markets. The VECM is also used extensively in studying the prices in the supply chain between energy prices and agricultural commodity prices (Cabrera and Schulz, 2016; Nazlioglu and Soytas, 2012).

Another strand of related literature is the application of clustering in price analysis. The development of machine learning motivates studying price dynamics with time series clustering. K-means clustering and hierarchical clustering are common clustering methods. K-means clusters time series with randomly specified centroids to iteratively cluster time series based on their similarity to a pre-specified number of clusters. Hierarchical clustering creates a hierarchy of clusters based on the similarity of time series. The desired number of clusters can be chosen with the degree of dissimilarity among clusters. The correct specification on similarity and choice of methods are required when dealing with time series (Graskemper, Yu, and Feil, 2021).

In the current literature, there is no conclusion as to whether K-means or hierarchical clustering is superior to the other. For a small number of clusters, we know that K-means is more efficient. Hierarchical clustering requires more computation power, though it can reveal the nested structure among clusters. Chen et al. Chen and Rehman (2021) applied both K-means and hierarchical clustering to identify critical periods of elevated volatility in volatility in energy markets. Liu et al. Liu et al. (2023) studied the egg market in China using K-means clustering with DTW as a distance measure. Ultimately, though, the choice between K-means clustering and hierarchical methods comes down to the judgement of the researcher. I will argue that in our case study of ASF in Chinese hog markets, a data-driven approach to determining the number of clusters tips the balance to hierarchical is advantageous.

3 Institutional Background

China is the world's biggest producer and consumer of pork, and pork is the major source of protein in China. The average per capita pork consumption is 33.6 kg in 2021. China only imports a small portion of pork, so the hogs raised domestically are the supply to processors and consumers. Historically, the hog cycle in China was around 3 to 3.5 years (Chavas and Pan, 2020) and the majority of hogs in the market were raised in farm households on a relatively small scale.

ASF is classified by the World Organisation for Animal Health as a List A disease. The mortality of infected hogs is almost 100% (Costard et al., 2009), and the outbreak of ASF in August 2018 killed millions of hogs. The outbreak was especially difficult to control in China due to the absence of vaccines and limited sanitary practices to prevent the spread of the disease. As a result, the hog stock decreased significantly.

Beginning with the first reported case in August 2018 in Liaoning province, the Chinese government recognized the urgent need for comprehensive measures to contain the epidemic and minimize its economic ramifications. In response to the ASF outbreak, the Chinese government enacted a series of policy interventions aimed at controlling the spread of the disease. Quarantine protocols were rigorously enforced to isolate infected areas and prevent the further dissemination of ASF to unaffected regions. This involved the establishment of containment zones where infected farms were identified and isolated, with strict biosecurity measures implemented to limit the movement of live hogs. Shipping restrictions including a ban on inter-province shipping of live hogs were imposed along with regulating the transportation of piglets and pork products between regions to prevent the inadvertent spread of ASF through the movement of infected hogs or contaminated products. The primary strategy to eliminate infected animals and prevent further transmission within and across herds is culling. Infected and at-risk piglets, hogs, and sows were culled on a large scale.

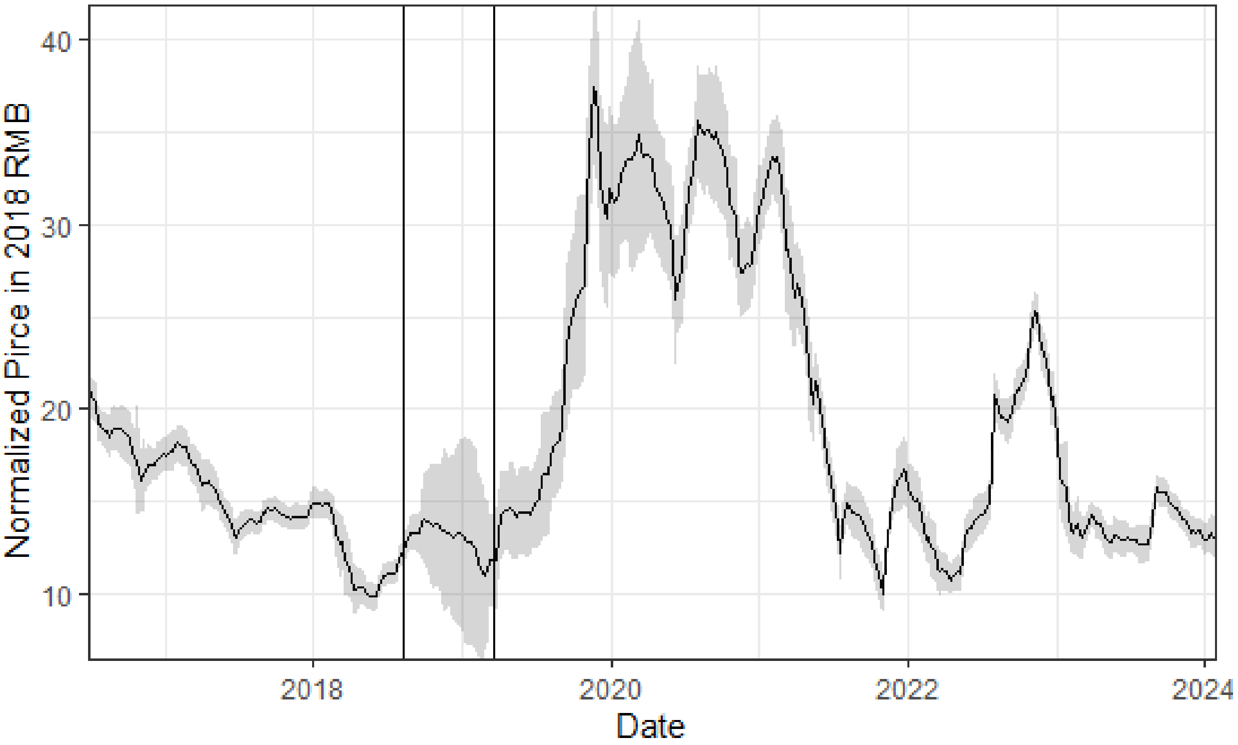
The Chinese hog market changed as a result of the inter-province shipping ban on live hogs as well as the large volume of culling during the outbreak of African Swine Fever (ASF) (Wang, Chavas, and Li, 2023). First, the culling of hogs and piglets due to the ASF caused a drastic drop in inventory. A breeding sow will have on average two litters per year. For a piglet to mature to slaughter weight, it takes 5.5-6.5 months. The length of biological cycles in pork production made instantaneous refilling of the hog inventory impossible. Second, the inter-province shipping ban on live hogs forced the market to deviate from its original equilibrium. A ban on shipping was implemented as soon as a province reported ASF.

The disturbance of market equilibrium reset temporal dynamics in hog prices (Ma, Delgado, and Wang, 2024). Starting in Liaoning province in August 2018, as more provinces discovered ASF, the number of provinces with shipping bans increased. When a shipping ban is effective for a province, live hogs cannot be shipped out from that province. Before the last shipping ban was lifted in April 2019, all provinces in China were affected by ASF. Throughout ASF, the patterns of live hog trading were completely changed due to shipping bans. Third, the small-sized farms were the hardest hit during the ASF. New policies focusing on subsiding large producers were carried out while the market recovered from the ASF (Ma et al., 2021). As a result, a considerable number of small-scale producers exited the market making the Chinese hog market more concentrated.

4 Data

Our weekly province-level hog price data ranges from June 2016 to December 2023, a total of 366 weeks. The original data are daily and county-specific. We aggregate the data to the province-week level by simple averaging due to missing observations and low variation in daily prices.

The weekly data is plotted in Figure 1. From left to right, the vertical lines indicate the start of the ASF shipping ban and the removal of the ban on August 31, 2018, and March 19, 2019, respectively. The shaded band around the price curve represents the two-standard deviation interval around the national average price captured by the solid line. Each point on the upper (lower) bound of the band represents the average price plus (minus) two standard deviations. There are 27 Chinese provinces in the sample. Out of all the mainland provinces (31 total), we dropped Hainan, Ningxia, Qinghai, and Tibet due to large numbers of missing observations during the period of interest. Given that they are all small producers of hogs and small consumers of pork, we are not concerned about excluding them from the empirical analysis.



Note: The two vertical lines indicate the start of the ASF shipping ban on August 31, 2018, and the removal of the ban on March 19, 2019, respectively. The line is the average price over all provinces with the shaded region representing two standard deviations around the mean.

Figure 1: Weekly National Average of Chinese hog price

During the outbreak of ASF, we observe a significant increase in the variance of price. Recall that this was the time when inter-province shipping bans were implemented. This increase in price level could be due to the culling of infected hogs. Although the government released its frozen pork reserve during the ASF outbreak, the herd size decreased significantly causing the lag in price increases. The variance of hog price went down with the increase in piece level. After the price peaked in late 2019, the variance increased again. There is no clear price pattern after the ASF. Table 1 shows that the price after the ASF period was significantly higher along with a substantial increase in volatility.

	Mean	Std. Dev.	Min	Median	Max	Unit	Missing %
Pre ASF							
Province hog price	14.95	2.75	9.85	14.59	21.37	RMB/kg	1.63
Post ASF							
Province hog price	20.68	8.12	10.09	16.58	37.39	RMB/kg	1.77

Table 1: Summary Statistics

The period of high prices in level coincides with the outbreak of Covid-19, but we conclude that Covid has a limited impact on hog prices. The number of cities that were locked down was a lot less than the number of cities affected by the inter-province shipping ban on live hogs. The lockdown during Covid prevented people from moving but not food so the demand for hogs was not greatly affected. The ingredients of foods dining at home are similar to dining away from home. More importantly, Wuhan was the first city under lockdown in Jan. 2020, but we can observe that the price already peaked before the first lockdown in January 2020.

5 Empirical Model

This is a classical representation of VAR and VECM adopted from Rapsomanikis et al. (2006). The price series used in the model can be specified as $\mathbf{Y}_t = (y_{t,1}, \ldots, y_{t,m})'$, which is a $m \times 1$ vector. The VAR model of order p, i.e. $\mathbf{VAR}(p)$ is:

$$\mathbf{Y}_t = \mathbf{c} + \mathbf{\Phi}_1 \mathbf{Y}_{t-1} + \mathbf{\Phi}_2 \mathbf{Y}_{t-2} + \cdots + \mathbf{\Phi}_p \mathbf{Y}_{t-p} + \boldsymbol{\epsilon}_t, \tag{1}$$

where $\mathbf{c} = (c_1, \ldots, c_m)'$ is a vector of intercept terms, the $\mathbf{\Phi}_k$ are $m \times m$ matrices of coefficients, and $\boldsymbol{\epsilon}_t = (\epsilon_{t,1}, \ldots, \epsilon_{t,m})'$ with the $\boldsymbol{\epsilon_t}$ being multivariate $\mathbf{iid}\left(\mathbf{0_m}, \boldsymbol{\sigma^2}\mathbf{I_m}\right)$. If the price series $\mathbf{Y}_t$ is $I(0)$, the VAR model is stable. Usually, the price series $\mathbf{Y}_t$ follows $I(1)$ then the first difference is required to transform the series to $I(0)$.

When the price series $\mathbf{Y}_t$ is $I\left(1\right)$ and cointegrated, the cointegrated VAR, i.e., the VECM can be expressed as follows with $\Delta\mathbf{Y}_t$ being the first difference of $\mathbf{Y}_t$ (Engsted and Johansen, 1997).

$$\Delta\mathbf{Y}_t = \mathbf{c} + \mathbf{\Pi}\mathbf{Y}_{t-1} + \mathbf{\Gamma}_1\Delta\mathbf{Y}_{t-1} + \mathbf{\Gamma}_2\Delta\mathbf{Y}_{t-2} + \cdots + \mathbf{\Gamma}_p\Delta\mathbf{Y}_{t-(p-1)} + \boldsymbol{\epsilon}_t, \tag{2}$$

where $\mathbf{\Pi} = \left(\mathbf{\Phi}_1 + \mathbf{\Phi}_2 + \ldots \mathbf{\Phi}_p - \mathbf{I}_m\right)$ and $\mathbf{\Gamma}_i = \left(-\sum_{j=i+1}^{p}\mathbf{\Phi}_j\right)$. Based on Johansen's specification of the VECM (Johansen, 1988), $\mathbf{\Pi}$ has rank r, $0 < r < m$. If the rank of $\mathbf{\Pi}$ is 0, the cointegration assumption is violated; if the rank of $\mathbf{\Pi}$ is m, then the system is stationary i.e $\mathbf{Y}_t$ is $I\left(0\right)$. The $\mathbf{\Pi}$ matrix can be factored as $\mathbf{\Pi} = \boldsymbol{\alpha\beta'}$, where $\boldsymbol{\alpha}$ is a $m \times r$ matrix and $\boldsymbol{\beta'}$ is a $r \times m$ matrix.

The factorization of the $\mathbf{\Pi}$ matrix is based on its rank. The rank r that maximizes the likelihood is the number of cointegrations. The $\boldsymbol{\beta}$ matrices can be interpreted as cointegrating vectors for the long-run equilibrium. The linear combinations of the series in $\mathbf{Y_t}$ that is stationary then the cointegrated series have mean-reverting behavior. The $\boldsymbol{\alpha}$ can be interpreted as the speed of adjustment parameters determining the speed of response to get back to the long-run equilibrium after a shock.

The VAR model contains a system of equations and each variable gets a coefficient. The number of coefficients grows quadratically even with a **VAR**(1) model as the number of series grows. In high dimensions, estimating such a model will have an over-fitting issue. LASSO estimation could prevent over-fitting, but it lacks the capability of statistical inference.

Apart from the difficulty in estimating numerous coefficients, the stationarity assumption of $\mathbf{\Pi}\mathbf{Y}_{t-1}$ is strong in the VECM. As the dimension of data grows, the potential linear combinations of series that are stationary grow as well. In such cases, estimation using maximum likelihood estimation will result in a flat likelihood function where getting the maxima is challenging.

5.1 Dimension Reduction

In order to fit a VECM that allows us to do statistical inference based on high-dimensional data, we need to reduce the dimensionality of the data. In practice, the number of series in a system should not exceed 11 (Pfaff, 2008). The reason for limiting the maximum number of series is to maintain the asymptotic property for the likelihood ratio test involving the number of cointegrating equilibrium, the rank r of the $\mathbf{\Pi}$ matrix.

To effectively reduce the dimension of the data, we apply hierarchical agglomerative clustering to reduce dimension while preserving the most important features of the series. Hierarchical agglomerative clustering allows us to cluster series based on their similarity. This method is optimal for the following reasons: it produces a hierarchy of provinces revealing the structure; there is no need to pre-specify the number of clusters; it is deterministic which guarantees the result of the co-integration test is also deterministic after dimension reduction. Since in our case, we cannot infer from industrial knowledge the optimal number of clusters, it is not ideal to use K-means clustering.

Compared to other methods like geographical clustering which collapses several price series into one (Wang, Chavas, and Li, 2023), hierarchical agglomerative clustering takes into account the price similarity in series. Price series in one cluster are similar enough that we can extract the centroid of each cluster to represent the cluster and therefore reduce the dimension. After dimension reduction, we use the VECM with Johansen's specification on the centroid of the clusters i.e. the mean of all time series of a cluster. We choose the mean as it minimizes the squared distance among series.

In clustering hog price data, the procedure begins by treating each time series as its own cluster. Initially, each series is assigned to a separate cluster. At each iteration, the two closest clusters, based on the DTW distance between their series, are merged. The closeness of clusters is determined by the Ward linkage method, which minimizes the increase in the total within-cluster variance when two clusters are combined. This process continues iteratively, merging the closest clusters until all time series are grouped into a single cluster.

The number of clusters is determined by at which iteration the process stops.

We use DTW to define distance and produce the distance matrix for the benefits of allowing for the alignment of time series that may have varying lengths or temporal distortions. This feature is particular desirable for price as price transmission takes a short time. The general form of the dynamic programming algorithm searching for the alignment with the minimal sum of distance between any two series Montero and Vilar (2015):

$$d_{\mathrm{DTW}}(\mathbf{X}_T, \mathbf{Y}_T) = \min_{r \in M} \left(\sum_{i=1}^{m} \|\mathbf{X}_{a_i} - \mathbf{Y}_{b_i}\| \right) \tag{3}$$

where M is the set of all possible sequences of m pairs preserving the order of the observations in the form $r = ((\mathbf{X}_{a_1}, \mathbf{Y}_{b_1}), \ldots, (\mathbf{X}_{a_m}, \mathbf{Y}_{b_m}))$ with $a_i, b_j \in \{1, \ldots, T\}$ such that $a_1 = b_1 = 1, a_m = b_m = T,$ and $a_{i+1} = a_i$ or $a_i + 1$ and $b_{i+1} = b_i$ or $b_i + 1,$ for $i \in \{1, \ldots, m-1\}$.

We choose Ward distance as linkage between clusters which can be viewed as the ANOVA sum of squares between the two clusters added up over all the series. The mathematical expression is as follows:

$$d_{\mathrm{Ward}}(A, B) = \sum_{i \in A \cup B} \|\mathbf{X_i} - c_{A \cup B}\| - \sum_{i \in A} \|\mathbf{X_i} - c_A\| - \sum_{i \in B} \|\mathbf{X_i} - c_B\| \tag{4}$$

where $\mathbf{X_i}$ is the series and c_j is the centroid of the cluster. Ward distance ensures that at each level in the hierarchy, the within-cluster sum of squares (WCSS) is minimized over all partitions obtainable by merging two clusters from the lower hierarchy. The Ward method aims at the lowest growth in variance as the cluster merges.

After having the hierarchy of clusters, we determine the optimal number of clusters using Hubert statistic (Hubert and Arabie, 1985). This approach permits a good balance between number of clusters and efficiency. Optimal number of clusters is selected when the Hubert statistic gains the largest increase as the return for having additional cluster starts to diminish.

To show the advantage of time series clustering, we use CH index (Caliński and

Harabasz, 1974) to measure the quality of clusters. The CH index measures the between cluster sum of squares over the within cluster sum of squares with adjustment for number of clusters similar to adjusted R^2.

6 Results and Discussion

6.1 The Clusters

Time series clustering using Hubert statistic specifies 4 clusters before the outbreak of ASF and 5 clusters after. Geographical clustering which does not change with time has 6 clusters throughout the period of study. Compared with geographical clustering, time series clustering tends to have more unbalanced clusters meaning that the number of provinces within a cluster varies a lot. The number of provinces within a cluster for geographical clusters is between 3 and 7 whereas in time series clusters the number is between 1 and 10.

With less clusters specified for both before and after the outbreak of ASF, time series clusters have higher quality over geographical clusters. The CH Index for the pre-ASF period is 146.60 and 93.72 for time series and geographical clusters. In the post-ASF period, the CH Index is 628.33 and 520.38 for time series and geographical clusters. Higher CH indices indicate better clustering with fewer within-cluster variations and larger between-cluster variations. The number of observations is different for the pre and post-ASF periods, so we can only compare the CH Index of the same period.

The dendrogram 2 and 3 provide a visual representation of the clusters formed by the hierarchical time series clustering. It resembles a tree-like structure where each cluster is represented by a branch. The height of each branch in the dendrogram represents the distance (dissimilarity) between clusters. We cut the dendrogram according to the optimal number of clusters determined by the Huber statistic. Different colors are used to show the clusters.

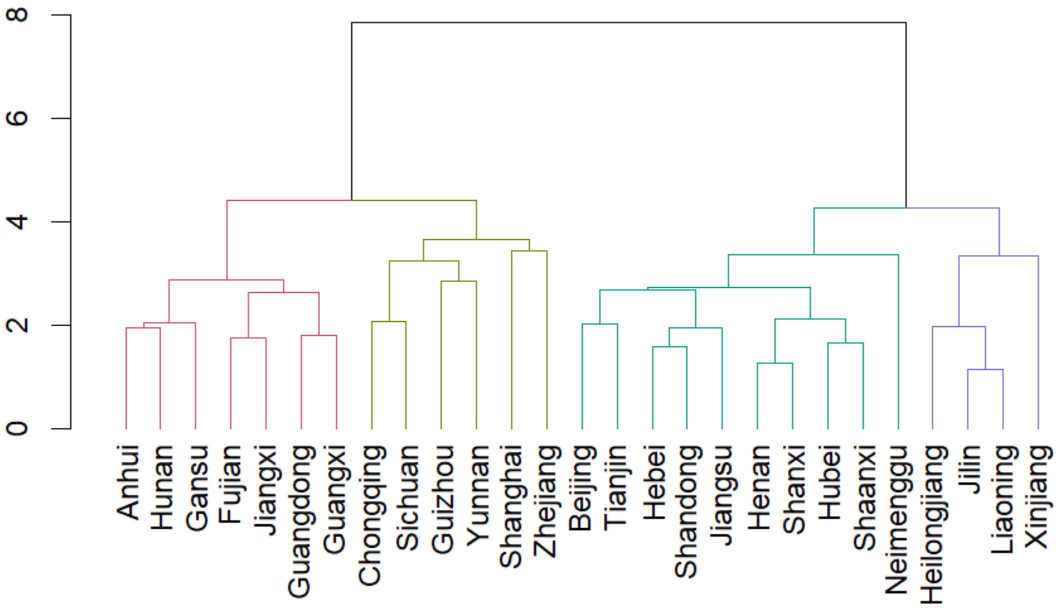


Figure 2: Dendrogram Pre-ASF

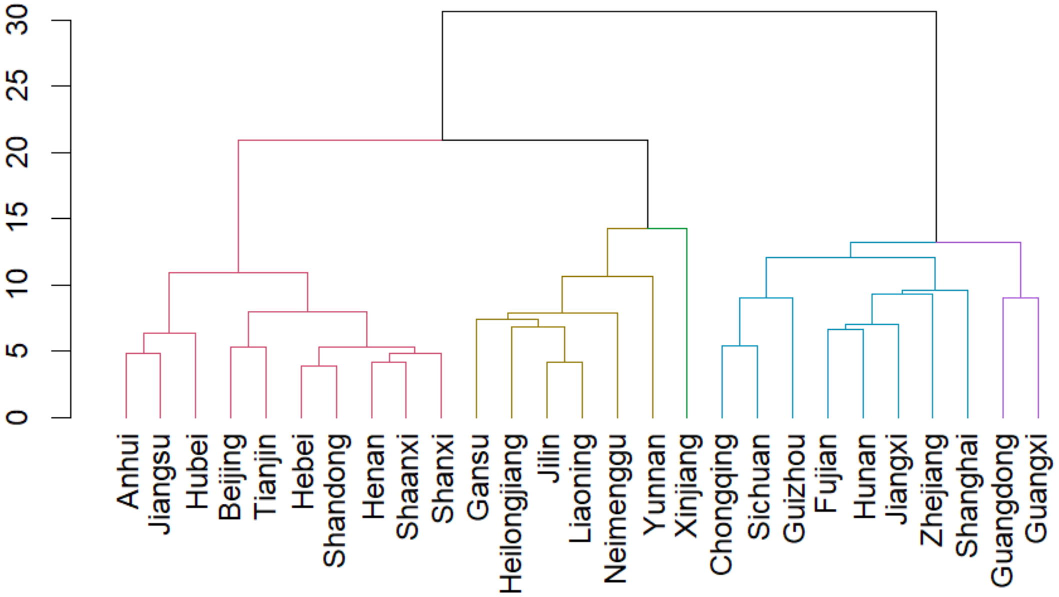


Figure 3: Dendrogram Post-ASF

In the pre-ASF period, the time series clusters show stronger spatial patterns with similarity to geographic clusters. Typically, China is divided into northern and southern regions by geographical differences. Two clusters contain provinces that are in both north and south regions, cluster 1 and 2. As shown in figure 4, cluster 2 with yellow color consists of Jiangsu which is a southern province while all other provinces are either in the middle or the north region. However, this result complies with our industrial knowledge. Shandong and Jiangsu are the most developed provinces in China in terms of total GDP. The transportation between the two provinces is also highly developed. Cluster 4 consists of provinces that are further apart. This result could be the fact that all provinces in cluster 4 are exporters of hogs. These provinces have larger hog production relative to their own consumption. The time series clustering result for the pre-ASF period is close to the grouping provided by the geographical clusters.

In the post-ASF period, the time series clusters are very different from the geographical clusters. We observe provinces in the north and south as well as in the east and southwest ending in the same cluster. From figure 5, we observe the clusters are distributed along latitudes. These patterns mimic the distribution of highways in China. Cluster 5 consists of only one province, Xinjiang. Similarly, cluster 4 consists of two provinces. These patterns could suggest less frequent and less random spatial arbitrages resulting in more local markets.

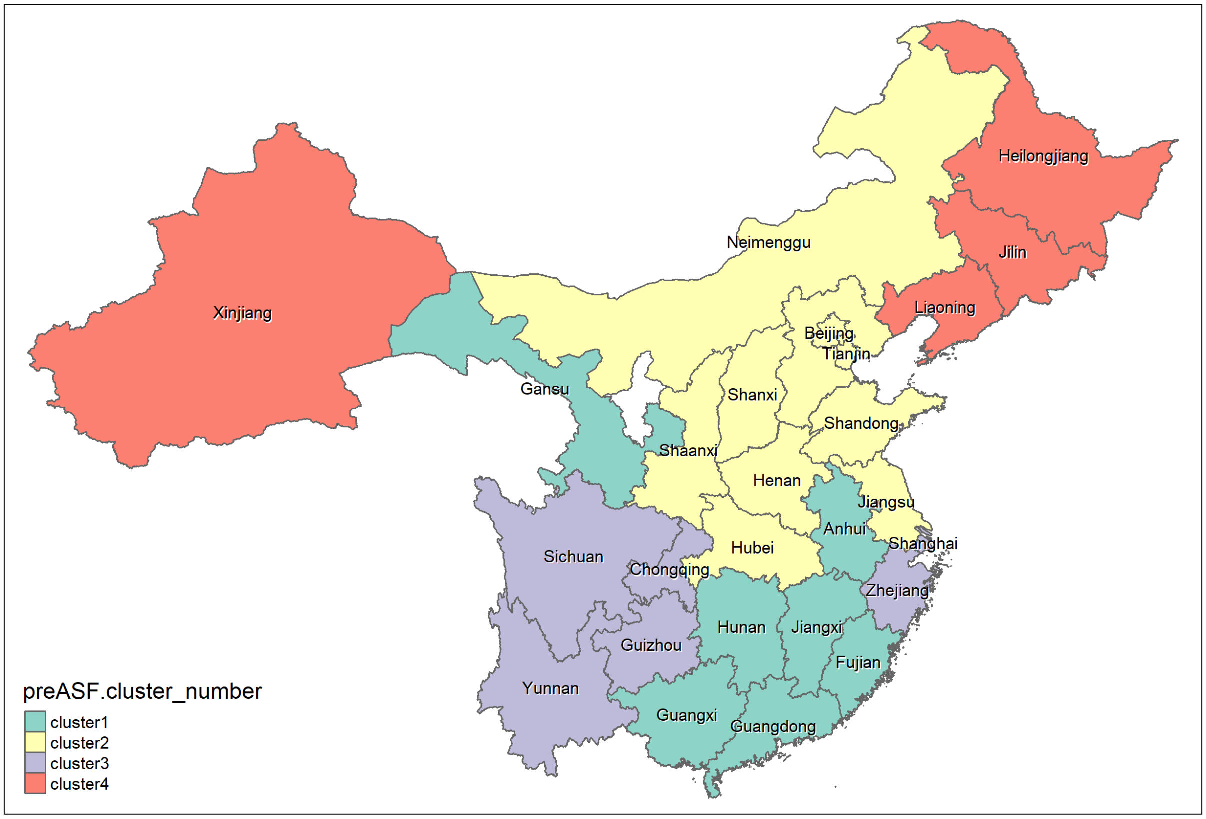
Cluster	Provinces
Pre ASF	
1	Anhui, Fujian, Gansu, Guangdong, Guangxi, Hunan, Jiangxi
2	Beijing, Hebei, Henan, Hubei, Jiangsu, Neimenggu, Shandong, Shaanxi, Shanxi, Tianjin
3	Chongqing, Guizhou, Shanghai, Sichuan, Yunnan, Zhejiang
4	Heilongjiang, Jilin, Liaoning, Xinjiang
Post ASF	
1	Anhui, Beijing, Hebei, Henan, Hubei, Jiangsu, Shandong, Shaanxi, Shanxi, Tianjin
2	Chongqing, Fujian, Guizhou, Hunan, Jiangxi, Shanghai, Sichuan, Zhejiang
3	Gansu, Heilongjiang, Jilin, Liaoning, Neimenggu, Yunnan
4	Guangdong, Guangxi
5	Xinjiang

Legend:

- Northeast: Heilongjiang, Jilin, Liaoning
- North: Beijing, Tianjin, Hebei, Shanxi, Neimenggu
- East: Shanghai, Jiangsu, Zhejiang, Anhui, Fujian, Jiangxi, Shandong
- Central South: Henan, Hubei, Hunan, Guangdong, Guangxi
- Southwest: Chongqing, Sichuan, Guizhou, Yunnan
- Northwest: Shaanxi, Gansu, Xinjiang

Note: The table is colored to show the geographical region of the provinces. The color code is in the legend.

Table 2: Time Series Clusters Pre and Post-ASF



Note: The dropped provinces are not shown in the graph.

Figure 4: Time Series Clusters Pre-ASF

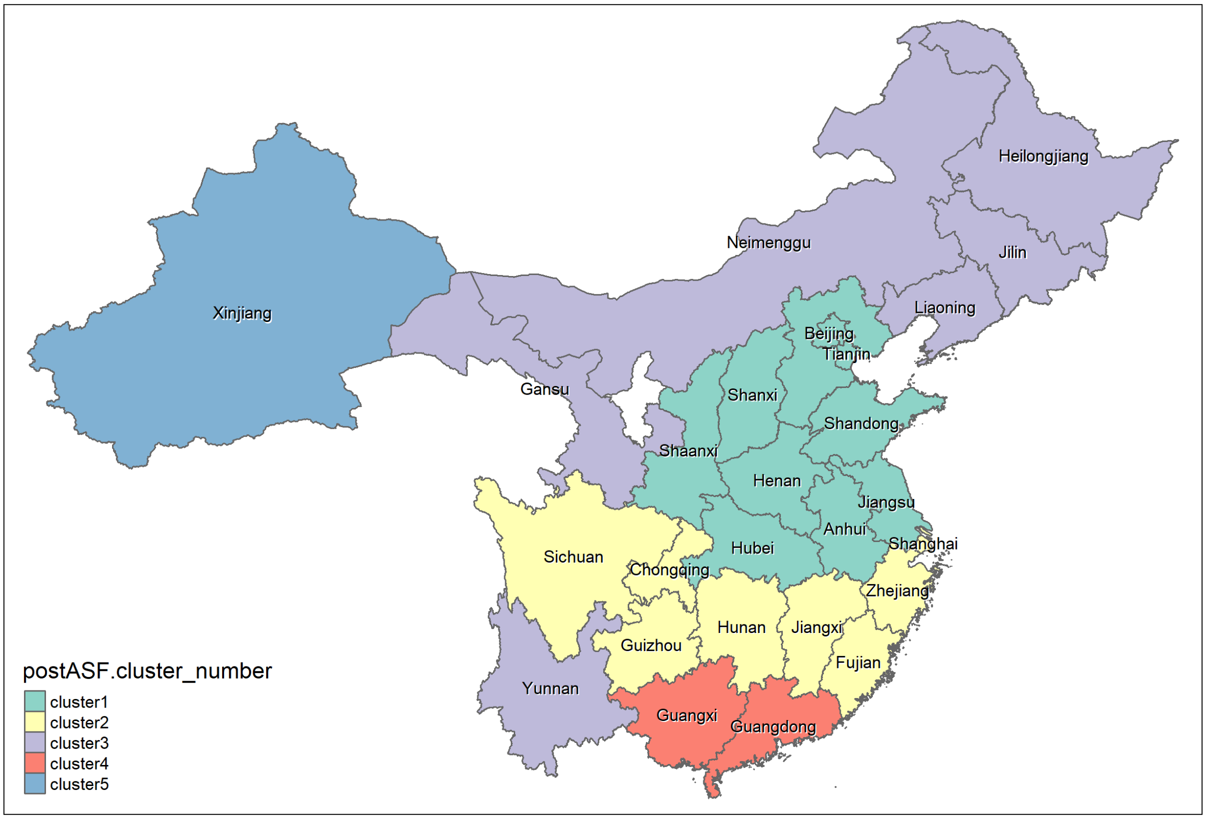


Figure 5: Time Series Clusters Post-ASF

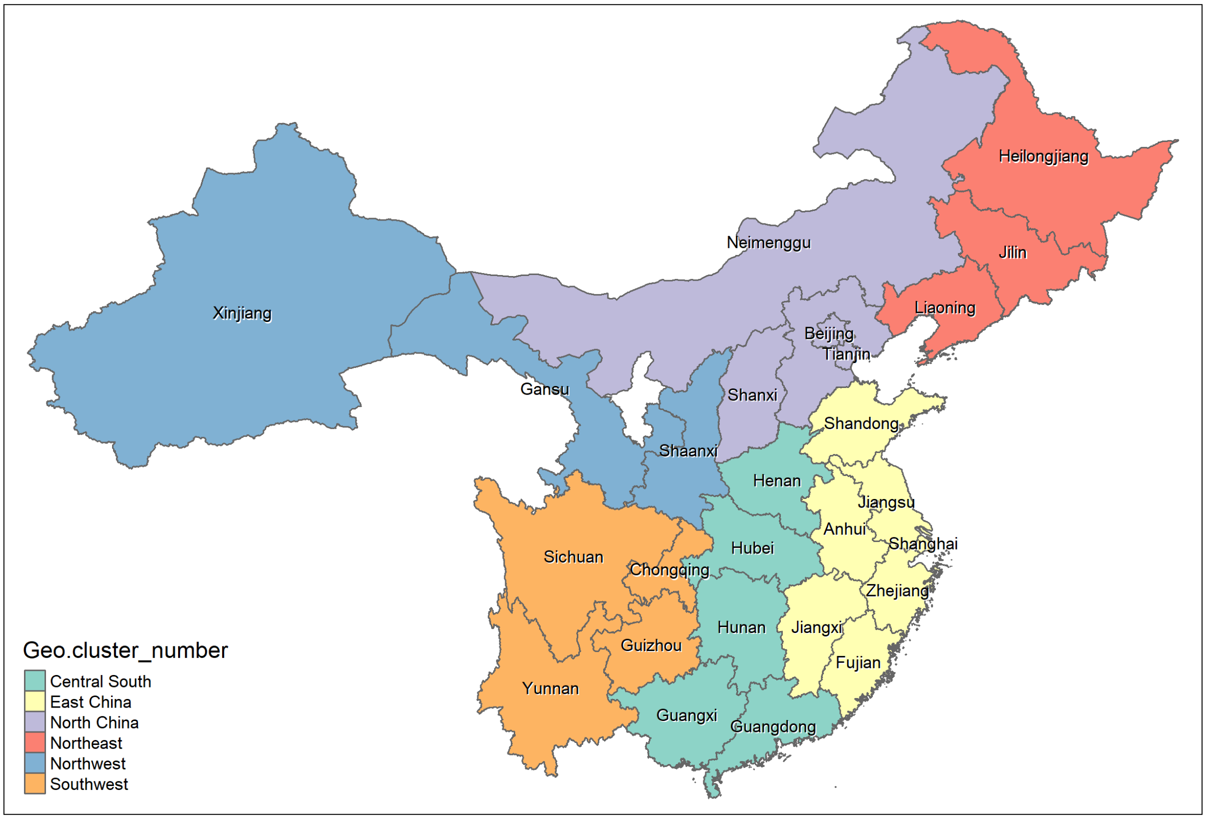


Figure 6: Geographical Clusters

6.2 VECM Results Among the Time Series Clusters

To investigate the potential changes in arbitraging patterns before and after the outbreak of ASF, we apply the VECM among the mean price of clusters for both the pre and post-ASF periods.

Augmented Dickey-Fuller test is used to confirm that all series are non-stationary during the pre-ASF and post-ASF period. The order of lag in the model is 1 which is selected by Akaike Information Criterion (AIC). The number of error correction terms (*ect*) of the cointegrating vector is determined by Johansen's cointegration test. We can connect the concept of cointegration with the mean reverting behavior. The cointegration vector is a linear combination of non-stationary series forming a stationary series. Since the cointegrating vector is stationary, its mean and variance do not change with time, so we can interpret the coefficients of *ect* terms as the speed of the mean reversion. The *cluster* in the regression table is the $\mathbf{\Gamma}$ in equation 2 showing the effect of own and cross lag of one period.

In the pre-ASF period shown in Table 3, *ect*1 is significant for all but cluster 3, and *ect*2 is significant for all clusters. Note that the coefficient of *ect*1 is negative and the coefficient of *ect*2 is positive for all clusters suggesting a strong mean reverting behavior.

Provinces in cluster 3 are relatively land-locked. The significant auto-regressive coefficient captures this fact. The easiest way to trade hogs between cluster 3 and the rest provinces is through Shannxi and Hubei which are in cluster 2. We observe that the coefficient for the lag of cluster 2 is significant in cluster 3's equation. Those observations on the significance of the coefficients provide evidence for the potential trading relationships between clusters.

The significant *ect* terms suggest strong mean reverting behavior as the result of active arbitrages. There is no data covering the trade flow of hogs between provinces, so recovering the potential trade patterns from the prices is the most practical approach.

The post-ASF period is much different from the pre-ASF period. The noticeable difference is that most of the *ect*s are no longer significant. In fact, the cointegrating vectors are not significant at all for cluster 2 and cluster 4. This result is suggesting that active and random arbitraging is no long present after the outbreak of ASF. The only significant coefficient for cluster 2 and 4 is the lag of cluster 3. We also observe that the lag of cluster 3 is significant for all clusters. This phenomenon could be caused by that cluster 3 was the first cluster to experience the ASF outbreak and therefore became the first cluster to declare ASF free and recover the herd size. It is very likely that other clusters are importing hogs from cluster 3 to rebuild their herd size.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
ect1	-0.574	***	-0.545	***	-0.146		-0.602	***
ect2	0.499	***	0.306	**	0.403	***	0.273	**
constant	0.108	***	0.099	***	0.030		0.108	***
cluster1	-0.357		-0.220		0.093		-0.201	
cluster2	0.680	*	0.358		0.770	***	0.199	
cluster3	-0.190		-0.070		-0.689	***	0.019	
cluster4	-0.116		0.120		-0.115		0.212	

Note: * indicates significance at the 0.05 level. Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Table 3: VECM of Time Series Clusters Pre-ASF

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
ect1	-0.433	*	-0.144		-0.354	*	0.061		-0.399	*
ect2	0.144		-0.116		0.140		-0.110		-0.196	
ect3	0.216		0.132		0.160		0.025		0.584	***
constant	0.025	*	0.007		0.021	.	-0.006		-0.002	
cluster1	-0.743	*	-0.442		-0.697	*	-0.350		-0.753	*
cluster2	0.186		0.021		0.135		-0.126		-0.201	
cluster3	1.108	***	0.915	***	1.081	***	0.816	**	1.578	***
cluster4	0.036		0.106		0.027		0.138		0.185	*
cluster5	-0.049		-0.063		-0.031		0.025		-0.227	.

Note: * indicates significance at the 0.05 level. Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Table 4: VECM of Time Series Clusters Post-ASF

6.3 VECM Results Among the Geographical Clusters

Unlike the time series clusters, the geographical clusters do not change in the number of clusters or the composition of a cluster. These clusters are constructed by taking the mean of provinces that belong to the same geographical regions specified by the Chinese government. The static nature of geographic clusters makes no distinctive patterns in cointegrating vectors before and after the outbreak of ASF.

The results from geographical clustering are very different from time series clustering in the pre-ASF period. The number of cointegration vectors is the same but we do not observe much significance of the *ect*s in the geographical clusters. Compared with time series clustering VECM results, geographical clustering does not preserve the information that allow us to find potential cointegrating equilibrium.

On the other hand, the post-ASF period shows some similarity in the time series cluster and geographical clusters. Both results show significance in the Northeast region (the Northeast region is contained in cluster 3) in all clusters' equation. This result suggests a similar argument that the Northeast region was the first to restore its herd size and became a price leader.

	NE		N		E		CS		SW		NW	
ect1	0.027		0.041		-0.067	.	0.026		-0.032		0.004	
ect2	-0.037		-0.071		0.261	*	0.013		0.197	.	0.056	
constant	-0.082		-0.104	*	-0.018		-0.139	**	-0.103	*	-0.091	*
NE	0.235		0.279		0.044		-0.050		-0.082		-0.005	
N	0.505		0.287		0.817	**	0.539	.	0.539	*	0.694	**
E	0.261		0.179		-0.489	.	0.384		0.177		0.193	
CS	0.261	**	-0.902	*	-0.338		-0.976	**	-0.335		-0.776	**
SW	0.065		0.010		-0.135		-0.029		-0.264		-0.132	
NW	0.480		0.551	.	0.294		0.459	.	0.136		0.466	*

Note: * indicates significance at the 0.05 level. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
NE: Northeast China. N: North China. E: East China. CS: Central South China. SW: South West China. NW: North West China.

Table 5: VECM of Geographical Clusters Pre-ASF

	NE		N		E		CS		SW		NW	
ect1	0.007		0.363		0.242		0.292		0.449		0.431	*
ect2	-0.089		-0.528		-0.168		-0.253		-0.472	.	-0.190	
ect3	-0.014		0.084		-0.157		-0.032		-0.121		-0.155	
ect4	0.069		0.125		0.221	*	0.136		0.322	**	0.169	
constant	0.003		0.024		0.007		0.002		0.015		0.015	
NE	0.786	***	-0.081		0.764	***	0.655	**	0.741	***	0.886	***
N	-0.081		-0.416		-0.060		0.024		-0.165		0.016	
E	-0.604	*	-0.508	*	-0.785	**	-0.710	**	-0.529	*	-0.759	**
CS	0.386		0.411		0.526		0.348		0.412		0.263	
SW	0.180		0.120		0.126		0.191		0.014		0.299	.
NW	-0.225		-0.170		-0.117		-0.075		-0.041		-0.261	

Note: * indicates significance at the 0.05 level. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
NE: Northeast China. N: North China. E: East China. CS: Central South China. SW: South West China. NW: North West China.

Table 6: VECM of Geographical Clusters Post-ASF

7 Concluding Remarks

From the time series clustering VECM result, we can observe the change in the Chinese hog market from the price perspective. The advantage of time series clustering is preserving the unobserved drivers of price including trade dynamics in data aggregation. The inter-province shipping ban reshaped the trade patterns since the outbreak of ASF. Even though we could not draw a conclusion from the number of *ect*s, their significance allows us to observe that the market is no longer strongly cointegrated.

In the time series clustering results, we can infer that the market was trading actively pre-ASF since the *ect*s are showing larger effects. In the post-ASF period, the most significant coefficients are the own and cross lags of the cluster 3 prices. The majority of significant coefficients of clusters change from *ect*s to the lag of cluster 3. From these results, We interpret that arbitrage became less frequent and less random after the outbreak of ASF.

When the market became less cointegrated after the ASF, we found similarities in results between time series and geographical clustering. Both results suggest that Northeast China hog prices drive the entire market. The nature that Northeast China is a hog exporter and the first to recover after ASF could explain this result.

The increased geographical distance between provinces in a cluster after the outbreak of ASF suggests that the producers may trade with fixed patterns and have long-term contracts. Rather than trade with intimate neighbors, producers are willing to trade with clients further away. From our knowledge, the shipping ban during the outbreak of ASF was public information, but such special measures often come with a sharp cutoff and a very short grace period. Producers will face significant losses if they do not acquire the information on special measures like shipping bans quickly. As most producers are risk-averse, it is natural to trade with partners from whom they know well and could acquire relevant information quickly to form long-term relationships. Such partners may not be geographically close to producers, so producers factor in the risk of policy change in shipping live hogs even after the government lifted all the bans.

The exogenous shocks of ASF and policies reshaped the market in price level and price dynamics. After all special measures were lifted, producers had more considerations of risks leading to different trading behavior. The price of live hogs in China did not return to its original level for years after the ASF. Both the results of VECM and clustering patterns show that price dynamics have yet to revert to the pre ASF patterns, and ultimately they may never return. Such findings encourage the government to consider the long-term effects of policies when implementing special measures during an abrupt outbreak of animal diseases.

We show that time series clustering effectively reduces data dimension while preserving the most distinctive information. In static geographical clustering results, it is hard to conclude changes before and after ASF. The comparison of time series clustering results before and after ASF suggests that trade and arbitrage are less active in the market.

One limitation of this case study is that no causality can be claimed from the time series clusters or the VECM. We hypothesize the potential reasons behind the VECM results and the patterns of the clusters before and after the outbreak of ASF. In the VECM, we can only have Granger causality. The time series clustering results suggest strong meaning reverting behavior in the price series potentially due to trade and arbitrage. We do not have ground truth about live hog trading across provinces, so we argue that the time series clustering accounts for trade based on our industrial knowledge.

References

Basu, S., and G. Michailidis. 2015. “Regularized estimation in sparse high-dimensional time series models.” *The Annals of Statistics* 43:1535 – 1567.

Beghin, J.C., and H. Schweizer. 2021. “Agricultural trade costs.” *Applied Economic Perspectives and Policy* 43:500–530.

Bressler, R.G., R.A. King, Jr., and R.A.K.R. Bressler. 1970. *Markets, prices, and interregional trade*, vol. 8. Wiley New York.

Cabrera, B.L., and F. Schulz. 2016. “Volatility linkages between energy and agricultural commodity prices.” *Energy Economics* 54:190–203.

Caliński, T., and J. Harabasz. 1974. “A dendrite method for cluster analysis.” *Communications in Statistics-theory and Methods* 3:1–27.

Chavas, J.P., and F. Pan. 2020. “The dynamics and volatility of prices in a vertical sector.” *American Journal of Agricultural Economics* 102:353–369.

Chen, J.M., and M.U. Rehman. 2021. “A pattern new in every moment: The temporal clustering of markets for crude oil, refined fuels, and other commodities.” *Energies* 14:6099.

Costard, S., B. Wieland, W. De Glanville, F. Jori, R. Rowlands, W. Vosloo, F. Roger, D.U. Pfeiffer, and L.K. Dixon. 2009. “African swine fever: how can global spread be prevented?” *Philosophical Transactions of the Royal Society B: Biological Sciences* 364:2683–2696.

De, Z., and D. Koemle. 2015. “Price transmission in hog and feed markets of China.” *Journal of Integrative Agriculture* 14:1122–1129.

Engle, R.F., and C.W. Granger. 1987. “Co-integration and error correction: representation, estimation, and testing.” *Econometrica: journal of the Econometric Society* 55:251–276.

Engsted, T., and S. Johansen. 1997. “Granger’s representation theorem and multicointegration.”, pp. .

Esposti, R., and G. Listorti. 2013. “Agricultural price transmission across space and commodities during price bubbles.” *Agricultural Economics* 44:125–139.

Gardner, B.L. 1975. “The farm-retail price spread in a competitive food industry.” *American Journal of Agricultural Economics* 57:399–409.

Graskemper, V., X. Yu, and J.H. Feil. 2021. “Farmer typology and implications for policy design–An unsupervised machine learning approach.” *Land Use Policy* 103:105328.

Hatzenbuehler, P.L., P.C. Abbott, and T. Abdoulaye. 2017. “Price transmission in Nigerian food security crop markets.” *Journal of Agricultural Economics* 68:143–163.

Hubert, L., and P. Arabie. 1985. “Comparing partitions.” *Journal of classification* 2:193–218.

Johansen, S. 1988. “Statistical analysis of cointegration vectors.” *Journal of economic dynamics and control* 12:231–254.

Liu, C., L. Zhou, L. Höschle, and X. Yu. 2023. “Food price dynamics and regional clusters: machine learning analysis of egg prices in China.” *China Agricultural Economic Review* 15:416–432.

Ma, M., M.S. Delgado, and H.H. Wang. 2024. “Risk, arbitrage, and spatial price relationships: Insights from China’s hog market under the African Swine Fever.” *Journal of Development Economics* 166:103200.

Ma, M., H.H. Wang, Y. Hua, F. Qin, and J. Yang. 2021. “African Swine Fever in China: Impacts, responses, and policy implications.” *Food Policy* 102:102065.

Mitra, S., and J.M. Boussard. 2012. “A simple model of endogenous agricultural commodity price fluctuations with storage.” *Agricultural economics* 43:1–15.

Montero, P., and J.A. Vilar. 2015. “TSclust: An R package for time series clustering.” *Journal of Statistical Software* 62:1–43.

Morrison Paul, C.J., and J.M. MacDonald. 2003. “Tracing the effects of agricultural commodity prices and food costs.” *American journal of agricultural economics* 85:633–646.

Nazlioglu, S., and U. Soytas. 2012. “Oil price, agricultural commodity prices, and the dollar: A panel cointegration and causality analysis.” *Energy Economics* 34:1098–1104.

Palaskas, T.B. 1995. “Statistical analysis of price transmission in the European Union.” *Journal of Agricultural Economics* 46:61–69.

Pfaff, B. 2008. *Analysis of integrated and cointegrated time series with R*. Springer Science & Business Media.

Rapsomanikis, G., D. Hallam, P. Conforti, et al. 2006. “Market integration and price transmission in selected food and cash crop markets of developing countries: review and applications.” *Agricultural commodity markets and trade*, pp. 187–217.

Roehner, B.M. 1996. “The role of transportation costs in the economics of commodity markets.” *American Journal of Agricultural Economics* 78:339–353.

Tokgoz, S., E. Wailes, E. Chavez, et al. 2011. “A quantitative analysis of trade policy responses to higher world agricultural commodity prices.” *Food Policy* 36:545–561.

Vachal, K. 2015. “Northern Plains Grain Farm Tuck Marketing Patterns.” In *Journal of the Transportation Research Forum*. vol. 54, pp. 85–98.

von Cramon-Taubadel, S., and B.K. Goodwin. 2021. “Price transmission in agricultural markets.” *Annual Review of Resource Economics* 13:65–84.

Wang, L., J.P. Chavas, and J. Li. 2023. “The dynamic impacts of disease outbreak on vertical and spatial markets: the case of African Swine Fever in China." *Applied Economics* 55:2005–2023.