Received 9 January 2022
Revised 9 May 2022
5 August 2022
Accepted 11 August 2022

Food price dynamics and regional clusters: machine learning analysis of egg prices in China

Chang Liu
College of Economics and Management, Jilin Agricultural University,
Changchun, China
Lin Zhou
Institute of Food and Nutrition Development,
Ministry of Agriculture and Rural Affairs of the People's Republic of China,
Beijing, China, and
Lisa Höschle and Xiaohua Yu
Department of Agricultural Economics and Rural Development,
Georg-August Universität Göttingen, Göttingen, Germany

Abstract

Purpose – The study uses machine learning techniques to cluster regional retail egg prices after 2000 in China. Furthermore, it combines machine learning results with econometric models to study determinants of cluster affiliation. Eggs are an inexpensiv, nutritious and sustainable animal food. Contextually, China is the largest country in the world in terms of both egg production and consumption. Regional clustering can help governments to imporve the precision of price policies and help producers make better investment decisions. The results are purely driven by data.

Design/methodology/approach – The study introduces dynamic time warping (DTW) algorithm which takes into account time series properties to analyze provincial egg prices in China. The results are compared with several other algorithms, such as TADPole. DTW is superior, though it is computationally expensive. After the clustering, a multinomial logit model is run to study the determinants of cluster affiliation.

Findings – The study identified three clusters. The first cluster including 12 provinces and the second cluster including 2 provinces are the main egg production provinces and their neighboring provinces in China. The third cluster is mainly egg importing regions. Clusters 1 and 2 have higher price volatility. The authors confirm that due to transaction costs, the importing areas may have less price volatility.

Practical implications – The machine learning techniques could help governments make more precise policies and help producers make better investment decisions.

Originality/value – This is the first paper to use machine learning techniques to cluster food prices. It also combines machine learning and econometric models to better study price dynamics.

Keywords Dynamic time warping (DTW), Machine learning, Egg price, Clustering, China

Paper type Research paper

1. Introduction

Ensuring price stability and food security is one of the top priorities of the policy agenda in many countries, and China, with the largest population in the world, is no exception (Yu *et al.*, 2020). Food is the primary need for human beings, and food price stability gets great attention both from the government and researchers (Tian and von Cramon-Taubadel, 2020).

The current literature demonstrates that drastic price fluctuations are often accompanied by a sudden shock outbreak, imbalances between supply and demand, loss of household welfare and even food crises and social panic (Bastianin *et al.*, 2016; Bellemare *et al.*, 2013;

The authors sincerely thank the two anonymous reviewers for the precious comments. The authors acknowledge the funding support from the National Natural Science Foundation of China (Fund ID: 71933004).

China Agricultural Economic Review
Vol. 15 No. 2, 2023
pp. 416-432
© Emerald Publishing Limited
1756-137X
DOI 10.1108/CAER-01-2022-0003

Sun *et al.*, 2017; Wang *et al.*, 2021b; Yu, 2014a). According to Bellemare *et al.* (2013) and Yu (2014b), the poor suffer more from the instabilities of food prices as they count with a higher budget share for food. Meanwhile, producers may also suffer from welfare loss due to dramatic price fluctuations (von Cramon-Taubadel and Goodwin, 2021). Food price stabilization, therefore, makes a substantial contribution to ensuring food security for the general public, as well as stable and healthy industry development in developing countries like China. In this case, appropriate policies to correct market failures are essential.

Many factors could affect food price volatility, such as energy prices, monetary policies, market speculations, natural disasters, air pollution and supply chain breakdowns (Sun *et al.*, 2017; Yu, 2014a, b; Yu *et al.*, 2020). After China's access to the WTO (World Trade Organization), China increased the import of food products, and domestic food price fluctuations began to be in tandem with the international market. Much instability in world food prices could stem from trade liberalization and government intervention (Tweeten, 1979, p. 231), and a local shock could be amplified to be a global crisis.

China is a large country, and regional heterogeneity of price volatility is assumed, however, the phenomenon is so far understudied. Identifying regional heterogeneities could help make more precise policies to intervene in price. The current literature often uses the club convergence test to study regional heterogeneities in time series (Tian *et al.*, 2016). However, it has many assumptions, such as data non-stationarity, and the results are not robust due to the core selection.

In contrast, we will use non-supervised machine learning techniques to identify the clusters of many time series. Such a technique is completely driven by the data and has very few assumptions. Specifically, provincial prices of egg products are our study object.

Eggs are a very important but inexpensive protein source for human beings. Eggs contain rich nutrients, including a high level of protein and various types of vitamins. Eggs are a great but also the lowest-cost source of animal-based protein (13.1 g per 100 g) and calories (139 kcal per 100 g), as well as the second-lowest-cost source of calcium (Rehault-Godbert *et al.*, 2019). Eggs are thus one of the most basic food products in China.

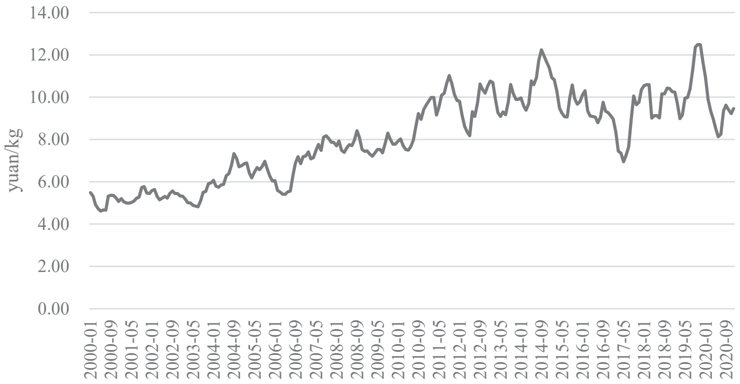
Eggs are also a very sustainable food. Laying hens can convert protein from their feed more efficiently than other animals. The protein feed conversion efficiency for eggs is 25%, for beef 3.8% and for pork 8.5% (Alexander *et al.*, 2016; Garnett, 2009). Overall, compared to other animal-based protein sources, egg production emits less CO_2 and requires less land per kilogram of protein (Herrero *et al.*, 2013; Nijdam *et al.*, 2012). This makes eggs an animal product with higher efficiency.

China is the largest egg producer and consumer in the world. According to the statistics of FAO (Food and Agriculture Organization of the United Nations), China produced 33.09 million tons of eggs (mainly chicken eggs) in 2019, accounting for 37% of the total global production. In the same year, China consumed a total of 29.84 million tons of eggs, with per capita consumption being 20.81 kg. Studying egg products can therefore help to make better nutrition and food security policies in China.

Figure 1 depicts monthly changes in egg prices in China after 2000. Clearly, we could observe an increasing trend of both price and price volatility.

Figure 1 shows that egg prices kept surging and reached a peak around 2010, and then maintained a high level of volatility. To some extent, price volatilities can be considered as a barometer for food crises (Yu *et al.*, 2020).

The current literature on price analysis mainly focuses on price transmission (von Cramon-Taubadel and Goodwin, 2021), including horizontal and vertical price transmission. However, to the best of our knowledge, little research has been devoted to the clustering of price series, especially taking the dynamic characteristic into account and employing appropriate methods tailored to time series. Clustering price dynamics can thus help identify different regional patterns and make precise policies.



Source(s): Ministry of Agriculture and Rural Affairs, China

Figure 1.
Monthly egg price in China after 2000

In light of this, we collected monthly egg prices for 30 provinces in China from the year 2000, and then used a non-supervised machine learning technique, specifically, the dynamic time warping (DTW) algorithm to cluster provincial prices. As a result, similar provinces in price patterns are clustered together.

We first review the existing literature both in models and algorithms in Section 2. Then Section 3 introduces our data set and all methodological algorithms. Section 4 exhibits applicable models of clustering in machine learning techniques. Section 5 compares the corresponding models introduced in Section 4. Section 6 discusses the results. Section 7 concludes our study and offers policy implications.

2. Literature review

Drastic price changes attracted increasing interests in economic theory and empirical studies. Traditional price prediction models, price transmissions and novel clustering methods in machine learning are applicable. Price prediction is one of the prevailing areas of price analysis. Researchers often employ ARIMA and ARDL models, two powerful forecasting techniques in price analysis in many areas (Dong *et al.*, 2020; Sun *et al.*, 2017). They can forecast future trends based on existing data. Jadhav *et al.* (2017) used the univariate ARIMA model to predict paddy, ragi and maize prices. To improve the prediction accuracy and select better models, many researchers combined ARIMA with machine learning techniques and achieved superior results (Dhini *et al.*, 2015; Kraemer *et al.*, 2020; Shahriar *et al.*, 2021). The introduction of machine learning has greatly enhanced the function of traditional models (Maruejols *et al.*, 2022) and facilitated access to unavailable data for analysis (Graskemper *et al.*, 2021, 2022; Wang *et al.*, 2021a).

The second strand of literature on price study is price transmission. According to the “one price law” and arbitrage, price transmissions occur between different regions or between links in the supply chain (Gardner, 1975; Zhou and Koemle, 2015). von Cramon-Taubadel and Goodwin (2021) have a comprehensive review of food price transmission.

With the increasing popularity of machine learning techniques more literature endeavors to adopt these techniques to study price dynamic patterns by identifying clusters. K-means and hierarchical clustering are traditional techniques, though they are often biased due to not considering time series properties (Graskemper *et al.*, 2021, 2022). For instance, Chen and Rehman (2021) employed five clustering methods including K-means and hierarchical clustering to identify critical periods in energy markets, analyze trades and understand temporal volatility in financial markets. Conradt *et al.* (2016) extended time series models to panel data for clustering and enhanced the predictive power by including exogenous variables, namely climate variable.

K-means and hierarchical clustering have limited power in clustering time series. On the one hand, the most common Euclidean distance adopted by K-means does not allow for time series shifts and is only applicable for series of equal length. However, time shifts and unequal lengths are potential features of time series. For instance, two price series might have the same volatility (same trend) because of a certain event but differ in the timing of their beginning response. This means that even series with the same volatility curves but different starting points cannot be identified as similar ones by K-means due to the way the distance is calculated. On the other hand, hierarchical clustering is usually appropriate for small data sets because of the inherent calculation procedures, whereas high dimensionality and large size are typical for time series. These reasons account for the deviation in the similarity calculation by K-means and hierarchical clustering methods.

However, DTW distance provides a solution to these concerns. It takes into account potential properties of dynamic time series, namely, unequal length, time shifts, high dimensionality and large size. By finding the optimal warping path of two time series, DTW then calculates the similarity by measuring the distance between the two series, though of different lengths. This means that even if one series is the subsequence of another time series, DTW can identify the similarity between the two. Hence, the superiority of DTW is mainly reflected in the handling of issues entailed by properties of dynamic data. With the help of DTW, more condensed clusters could be identified for high-dimensional time series. Various studies have already used DTW in gesture recognition (Wobbrock *et al.*, 2007), voice recognition (Ding and Shi, 2017) and finance (Tsinaslanidis, 2018).

DTW algorithm also shows some advantages in price clusters. By running a benchmark on a large number of time series, Pfisterer *et al.* (2019) found DTW is a reliable method in terms of time series clustering after assessing a range of methods performed on these data sets. Dmytrow *et al.* (2021) adopted DTW to measure similarities between price series in clustering energy commodities, and found a connection between COVID-19 and energy commodity prices. Yin and Shang (2014) combined DTW with DID (difference-in-differences) to create a randomized experiment to form a way to save electricity consumption. Similarly, Chai *et al.* (2019) proposed a hybrid model in which DTW was applied to group subsequences, and then ARIMA was introduced for the prediction of crude oil price. Nakagawa *et al.* (2019) used the DTW algorithms for stock price predictions.

However, to the best of our knowledge, little attention has been paid to clustering prices of agricultural products. Despite the declining share of agricultural GDP in China, agriculture still occupies a vital position in the national economy. In response, our paper particularly uses the DTW algorithm to cluster provincial egg prices to identify regional patterns.

3. Data description

Eggs are an animal product most widely consumed in China. We collected monthly averaged retail egg prices for 30 provinces (except for Tibet due to data unavailability) in China for the period from January 2000 to December 2020, from the Ministry of Agriculture and Rural Affairs of China (MARA). Our data set covers a period of 21 years, collected by MARA from local fresh markets. It represents the retail market price in China. Figure 1 shows the trend of the average price nationwide.

Egg production shows a regional concentration pattern. Figure 2 shows the provincial egg production in 2019. The top 4 provinces (Shandong, Henan, Hebei and Liaoning) produced 15.86 million tons, about 48% of total national production. The top 12 provinces, which have more than 1 million tons of eggs, produced 27.70 million tons, about 84% of the national production.

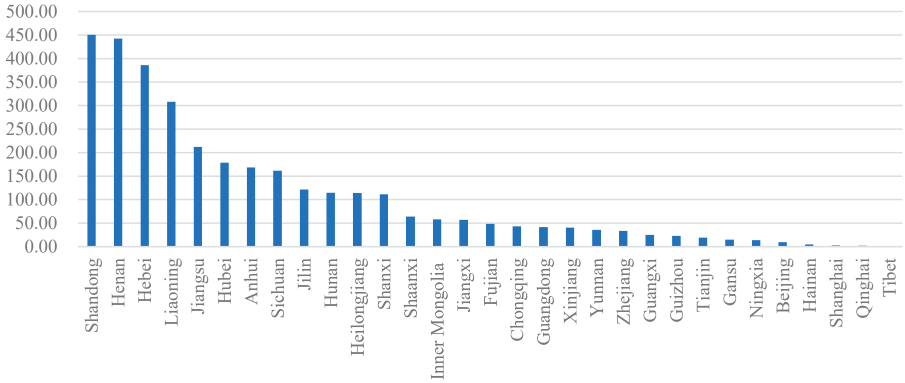


Figure 2.
Provincial egg production in 2019

Note(s): Unit: 10,000 ton

Source(s): China Rural Statistical Yearbook 2020. (Table 7-40, pp.187)

Here we define the 12 provinces with more than 1 million tons as the major production provinces (namely, Shandong, Henan, Hebei, Liaoning, Jiangsu, Hubei, Anhui, Sichuan, Jilin, Hunan, Heilongjiang, Shanxi). They are depicted in the map in Figure 3. They are located in Northeast China, North China and middle and lower reaches of the Yangtze River, and are traditionally plain areas with good agricultural production conditions.

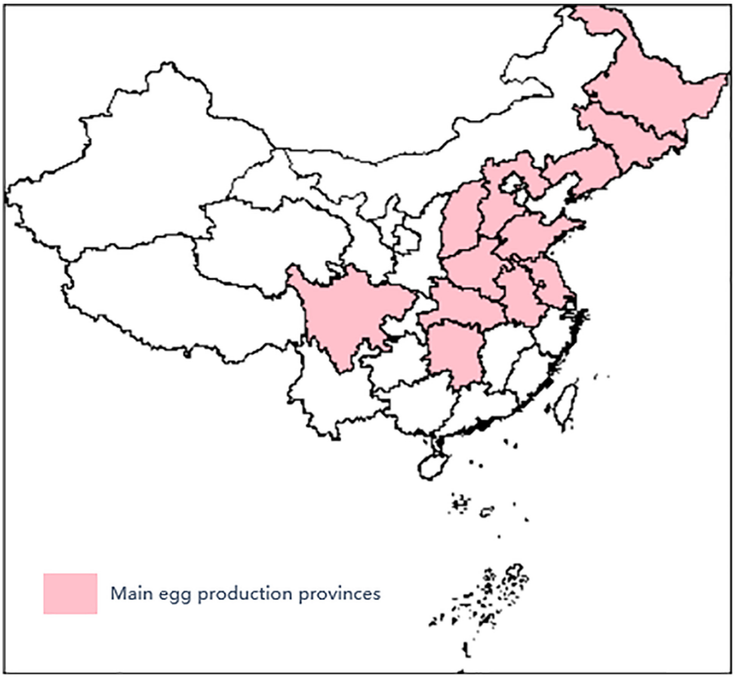


Figure 3.
Distribution of main egg production regions in China

Source(s): China Rural Statistical Yearbook 2020. (Table 7-40, pp.187)

4. Method

4.1 Dynamic time warping (DTW) distance

Distance or dissimilarity between observations is a fundamental measure for clustering. The traditional methods include K-means and hierarchical methods which, respectively, use the Euclidean distance and the shape-based distance (SBD) (Graskemper *et al.*, 2021, 2022; Sardá-Espinosa, 2019). However, the Euclidean distance does not fit time series properties well due to shift and unequal lengths of time series. Similarly, hierarchical clustering does not work well in terms of data size, inducing unintended bias to the clusters. Given these points, we will use the DTW distance as a dissimilarity measure to account for shifts in time series as well as high dimensionality properties (Aghabozorgi *et al.*, 2015). In this way, each series is assigned to an explicit cluster.

DTW is an algorithm commonly used to compare a pair of time series even if of different lengths or with time shifts. Therefore, DTW distance is tailored to measure the dynamic distance between one and another time series to determine an optimal curve (Berndt, 1994; Ratanamahatana and Keogh, 2004). The procedure of extracting centroid function is called DBA (DTW barycenter averaging).

For a series of different lengths, the DTW algorithm is often used to align two sequences by point-to-point mapping to measure the pairwise distance and find the minimized alignment by non-linear iterations. If we denote $X = (x_1, \ x_2, \ \ldots, \ x_m)$ and $Y = (y_1, \ y_2, \ \ldots, \ y_n)$ as a pair of sequences, then the first step in DTW is to create a local cost matrix $(m \times n)$ for every pair of sequences (i,j), $i = 1, 2, \ldots., n$ and $j = 1, 2, \ldots., m$,

$$lcm(i,j) = \left(\sum_v \left| x_i^v - y_j^v \right|^p \right)^{1/p} \tag{1}$$

where v is the set of all $(i, \ j)$ combinations, and p corresponds to the lp norm that was used to construct the local cost matrix (LCM).

In the second step, DTW sets out to calculate the optimal path by minimizing the alignment throughout LCM iteratively. Finally, after the determination of the optimal path in LCM, the DTW distance could then be calculated by Equation (2).

$$DTW_p(x,y) = \left(\sum \frac{m_{i,j} lcm(i,j)^p}{M_{i,j}} \right)^{1/p} \tag{2}$$

where $m_{i,j}$ is the weight of each step, and $M_{i,j}$ is a normalization constant (Giorgino, 2009).

Once we have the distance measure of DTW, we can extract the centroid function which is called DBA.

Different from conventional clustering methods (e.g. K-means and hierarchical clustering), partitional clustering with DTW assigns observations into an explicit cluster, and each object can only be attributed to one cluster. By partitional clustering, the maximized inter-group distance and the minimized intra-group distance can be obtained in an iterative way.

4.2 Three methods of partitional clustering with DTW

4.2.1 Partitional clustering based on DTW distance and DBA centroids. This is the algorithm that directly computes DTW distance and extracts DBA centroids. It is very straightforward but rather computationally expensive both in terms of time and memory given the fact that the distance measure would be conducted $m \times n$ times for alignment.

4.2.2 Partitional clustering based on DTW lower bounds and DBA centroids. In order to reduce computational complexity, Keogh and Ratanamahatana (2005) and Lemire (2009) proposed DTW lower bounds algorithms. However, the lower bounds method has its applicable conditions, which means that all the time series are required to equip an equal length in computation, and thus re-interpolation is indispensable. A consistent lp norm of calculation between lower bounds and DTW is also needed despite the equal length.

4.2.3 Partitional clustering based on TADPole. TADPole clustering works under a new clustering framework proposed by Begum *et al.* (2015), in which time series clustering is available by adopting DTW distance and a centroid chosen from existing data (Begum *et al.*, 2015). The algorithm functions in a way similar to PAM (partition around medoids) clustering, because the medoid or centroid is always an element from the originally collected data.

Despite the specific distance and centroid, a cutoff distance value must be defined so that the algorithm can prune unnecessary calculations and thus speed up clustering. Additionally, similar to DTW lower bounds, equal length of all series is required, and re-interpolation is needed as well for TADPole. In this respect, it works in a way similar to traditional hierarchical clustering. Thus, series with the nearest distance condense together, and a certain cluster is formed in this manner.

4.3 Computation procedure and optimal number of clusters

The workflow of time series partitional clustering is similar to K-means except that it utilizes the DTW algorithm to measure the distance between samples.

The detailed procedure is offered as follows. First, randomly choose j centroids (provincial time series), and thus j clusters are randomly formed. Each provincial time series would be assigned to the cluster closest to the individual after distance computation. Second, the prototype of each cluster is extraced, updating the cluster by recomputing the distance between each provincial time series and each updated prototype. Third, iterations are realized based on the previous step until no change occures, or a certain number of iterations has been reached.

Another notion must be added, that the expected clustering number in partitional clustering is supposed to be set beforehand, and each province would be assigned to only one cluster progressively. We ex ante set up different numbers of clusters and compare them by internal cluster validity indices (CVIs) to identify the optimal number of clusters. Sardá-Espinosa (2019) provided a comprehensive review and comparison of different CVIs.

5. Model comparison

5.1 Selection of number of clusters

The expected number of clusters is supposed to be set before partitional clustering. Since no ground truth is given, internal CVIs supply an appropriate way to evaluate cluster numbers, as the internal indexes only consider the partitioned data and try to define a measure of cluster purity (Sardá-Espinosa, 2019). Though the literature proposed many different CVIs for the selection of the number of partitional clusters, we will use the following four internal indices which maximize CVIs, namely, silhouette index, score function index, Calinski-Harabasz index and Dunn index (Sardá-Espinosa, 2019).

Table 1, reports the related CVIs for the partitional clustering based on DTW distance and DBA centroids with the number of clusters from 3 to 5. All indices indicate that 3 is the optimal number of clusters.

Table 1. Cluster number evaluated by internal CVIs

Indices	$k = 3$	$k = 4$	$k = 5$
Silhouette	0.1063	0.0563	0.0966
Score function (SF)	0.0000	0.0000	0.0000
Calinski-Harabasz (CH)	14.9867	9.5917	5.7331
Dunn (D)	0.4978	0.4732	0.4135

Source(s): From the authors' computation

5.2 Selection of the algorithms with validity index

After we determined the optimal number of clusters for 3, we can compare the results between (1) partitional clustering based on DTW, (2) partitional clustering based on DTW lower bounds and (3) partitional clustering based on TADPole. Meila (2003) suggested using the variation of information (VI) to evaluate different clustering algorithms. The smaller the VI value is, the better the algorithm is.

Table 2 reports the final VI values for each clustering algorithm demonstrated in partitional clustering. As expected, DTW and the lower bounds of DTW have the same VI value of 0.3832, smaller than TADPole. We believe DTW would have the best results as it has fewer pre-requirements, though it is computationally expensive. In addition, DTW lower bounds also has the same clustering results as the DTW. It shows the robustness of the DTW algorithm. The following discussion will be mainly based on the DTW results (as well DTW lower bounds results), while the results for TADPoles are used for robust check and comparison.

6. Result discussion

6.1 Geographical distribution of the clusters

The main results discussed here are based on the computation of DTW algorithm, as it has fewer assumptions and shows robustness from the above comparisons. Table 3 summarizes the corresponding outcome with 3 clusters, in which provinces closest in distance are clustered together. Meanwhile, the average price of each cluster is presented, through which we can identify to what extent cluster prices differentiate from each other.

Figure 4 also plots the results in the map. It is clear that the clusters show geographical concentration patterns. The first cluster covers Northeast China, North China plain and a few connected provinces, such as Jiangsu Province, Shaanxi Province and Chongqing City. These regions are main Chinese food production areas. The second cluster is composed of Hubei Province and Anhui Province, which are the two Central China agricultural provinces. The rest belong to the third cluster. If we compare the clusters with Figure 3, it seems that clusters 1 and 2 are highly overlapping with the major egg production provinces. We will have a detailed analysis in the next sections.

Clustering algorithm	DTW	DTW_LB	TADPole
VI	0.3832	0.3832	0.4405

Note(s): All algorithms have 3 clusters

Table 2. Comparison of different clustering algorithms with VI

Cluster	1	2	3
Members	Beijing, Tianjin, Hebei, Shanxi, Liaoning, Jilin, Heilongjiang, Jiangsu, Shandong, Henan, Chongqing, Shaanxi	Anhui, Hubei	Inner Mongolia, Shanghai, Zhejiang, Fujian, Jiangxi, Hunan, Guangdong, Guangxi, Hainan, Sichuan, Guizhou, Yunnan, Gansu, Qinghai, Ningxia, Xinjiang
Average price (yuan/kg)	7.04	7.95	8.85
Total number	12	2	16

Table 3. Cluster result from DTW partitional

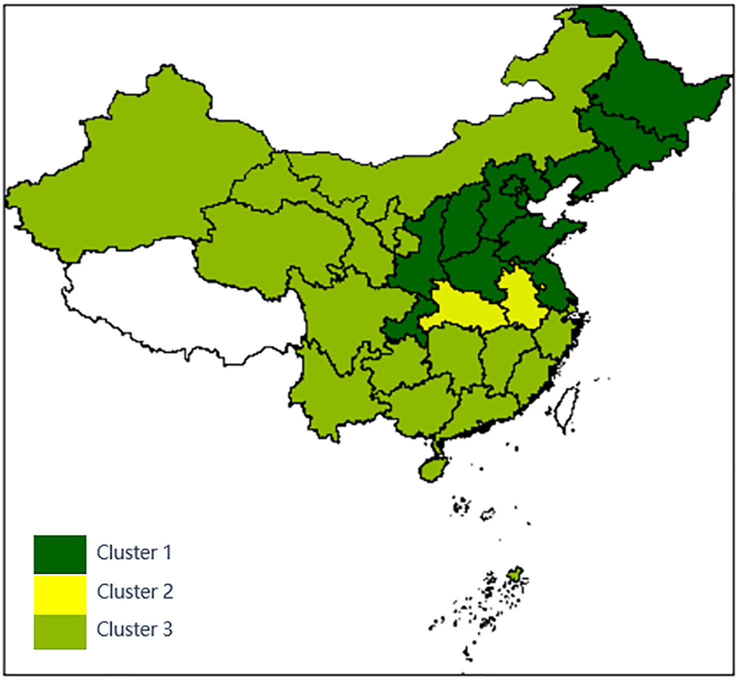


Figure 4.
The clusters of DTW algorithms

6.2 Price dynamics of the clusters

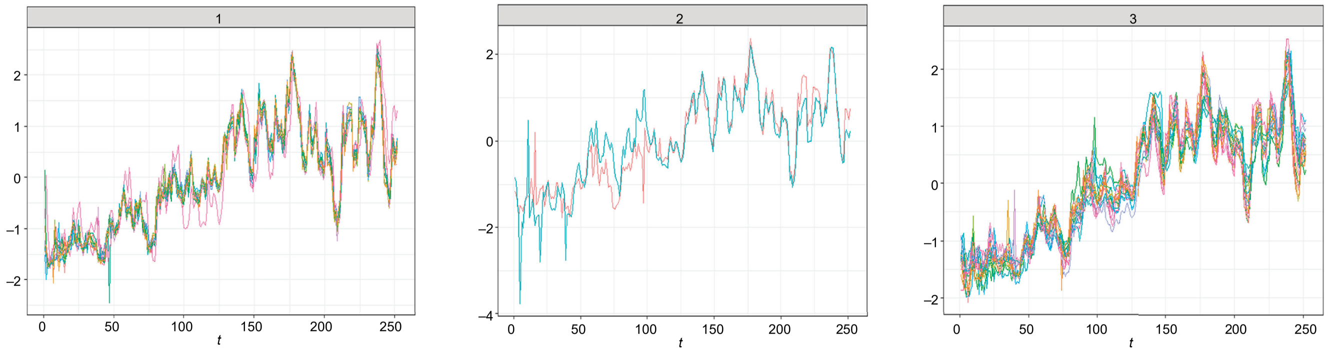
Figure 5 visualizes the price dynamics of the three clusters (Figure 5a–c). It supportes our clustering results, as prices within each cluster show similar patterns. We can observe that the three clusters demonstrate different levels of volatilities, particularly before 2010. Cluster 2 (Figure 5b) exhibits a high level of volatility, with both extremely high and extremely low prices in comparison with the other two clusters in the same period. In contrast, cluster 1 (Figure 5a) shows a medium level of volatility, and cluster 3 (Figure 5c) shows relatively lower price volatilities. A distinct observation is that cluster 3 has fewer extremely low egg prices.

6.3 Determinations of the clusters

We further run a multinomial logit to study the determinants of the clusters, as Section 6.1 shows a geographic pattern. The explanatory variables include per capita GDP, the share of agriculture in total GDP, land per farmer and the production of eggs (or whether a main egg production province [1 for yes, and 0 for no], or per capita egg production).

We only have 30 provinces in our clustering analysis. In order to have enough degrees of freedom, we cannot run a panel data model. Furthermore, the data cover the period from 2000 to 2020, and we use observations in 2010 as the dependent variables, which is the middle point of the sample period.

Table 4 shows the estimation results for the multinomial logit model which studies the determinants of the clusters. The coefficients for agricultural GDP share, and whether the province is a main egg production province (or the production quantity, or per capita production quantity) in cluster 1 (corresponding to results of models 1, 2 and 3, respectively) in comparison with cluster 3 (cluster 3 as the base) are statistically significant. However, no coefficients in cluster 2 results are statistically significant. The results are robust whether we



Cluster 1
(a)

Cluster 2
(b)

Cluster 3
(c)

Note(s): The numbers on the horizontal coordinate represent the nth month starting from January 2000. Since the observations span a total of 21 years from 2000 to 2020, there are 252 numbers on the horizontal coordinate, representing 252 months in 21 years. For example, the year 2010 corresponds to the period of 121-132 in the horizontal coordinate

Figure 5.
Price dynamics of three clusters

Table 4. Multinomial logit results for the determinants of the clusters (cluster 3 as the base)

	Cluster 1						Cluster 2					
	Model 1		Model 2		Model 3		Model 1		Model 2		Model 3	
	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value
Per capita GDP (10,000 Yuan)	−2.158	0.122	−0.885	0.392	−0.518	0.659	−2.951	0.326	−0.374	0.881	−0.300	0.924
Share of agriculture in Total GDP	−1.458	0.077*	−1.129	0.064*	−1.530	0.059*	−1.038	0.359	−0.037	0.960	−0.087	0.918
Average land size per farmer (mu per capita)	0.379	0.342	0.252	0.230	−0.340	0.258	0.072	0.918	−0.125	0.834	−0.527	0.462
Whether a main egg production province (1 for yes, and 0 for no)	7.412	0.049**					20.736	0.992				
Log (total egg production)			3.483	0.031**					2.214	0.123		
Log (egg production per capita)					9.085	0.037**					5.518	0.114
Constant	17.627	0.111	−1.342	0.859	−28.626	0.114	0.792	1.000	−10.1145	0.487	−27.610	0.238

Note(s): *, ** and *** denote 10%, 5% and 1% statistical significance, respectively

Note that it seems that inclusion of both per capita production and total production would cause multicollinearity problem in practice

Source(s): National Bureau of Statistics (2020), except that egg consumption data in 2019 are taken from the Statistical Year of China, 2020 include the production quantity or per capita production quantity or the dummy for major production regions with more than 1 million tons.

First, it implies that main egg production provinces are more likely to belong to cluster 1, which has relatively high price volatility. It confirms the rationale of overlapping provinces between the main production areas (depicted in Figure 3) and cluster 1 (depicted in Figure 4).

Cluster 1 includes a majority of egg-producing provinces and the 5 largest egg exporting provinces (Shandong, Henan, Hebei, Liaoning and Jiangsu). It can be understood that cluster 1 is in the inner circle of the egg-producing provinces, neighbored by Anhui and Hubei, two relatively large producers in cluster 2, then surrounded by provinces in cluster 3 at the outermost edge. That means, to which extent price fluctuation correlates with distance to egg-producing provinces. It is believed that provinces of remote distance to epicenters are less susceptible to the high frequency of volatilities due to transaction costs (Meyer and von Cramon-Taubadel, 2004). For instance, when there is an oversupply in production provinces, the price could plunge. However, the price in importing regions would not plunge too low due to transaction costs.

This can partly explain the differences that appeared in the three clusters, for the inter-provincial speculations largely rely on price gaps and distances, which takes time and costs for price transmission (von Cramon-Taubadel and Goodwin, 2021). Speculation only occurs when the price difference far exceeds the cost of the transaction. Therefore, for regions (cluster 3) remote from the epicenter, transaction costs increase with distance, say, a higher threshold exists for the provinces distant from large-scale egg production provinces, and they survive frequent price fluctuations because of a high transaction barrier to price transmission. This is also consistent with the average price of each cluster exhibited in Table 3. The further the areas (cluster 3) are from production bases, the higher the average price. However, despite the low frequency of volatilities in cluster 3 in the short term, the long-term fluctuations and upward trends largely remain consistent with the other two clusters, which indicates that the egg market in China is well integrated. This sheds light upon the common basis for domestic economic circulation (Yu *et al.*, 2021). Interestingly, the theory of price transmission has a thorough reflection on the clustering algorithms concerning time series.

Table 4 also shows that the coefficient for agricultural GDP share is negative and marginally significant. It shows that provinces with higher agricultural GDP share are less likely to be in the first cluster. It is possible that these major egg production provinces are mainly plains with large population, and have relatively good economic development levels. In these regions, the non-agricultural sector grows much faster than the agricultural sector.

6.4 Robust check

We use the TADPole clustering results as the robust check, which are visualized in Figure 6.

The results are substantially different from the DTW algorithm. Note that the VI index reported in Table 2 already indicated that TADPole results are not as good as DTW clustering.

Nonetheless, we find that cluster 1 in TADPole algorithm basically covers cluster 1 and cluster 2 in the DTW algorithm (also including Inner Mongolia and Hunan Province, as the extended region of the main production areas). In other words, cluster 1 in TADPole also overlaps the main egg production provinces. This is a consistent conclusion with the result from DTW.

6.5 Policy implications

Machine learning results reveal 3 clusters in terms of egg price dynamics in China. The first cluster of 12 provinces and the second cluster of 2 provinces are main egg production regions and have higher price volatility. The government of China should specifically target these

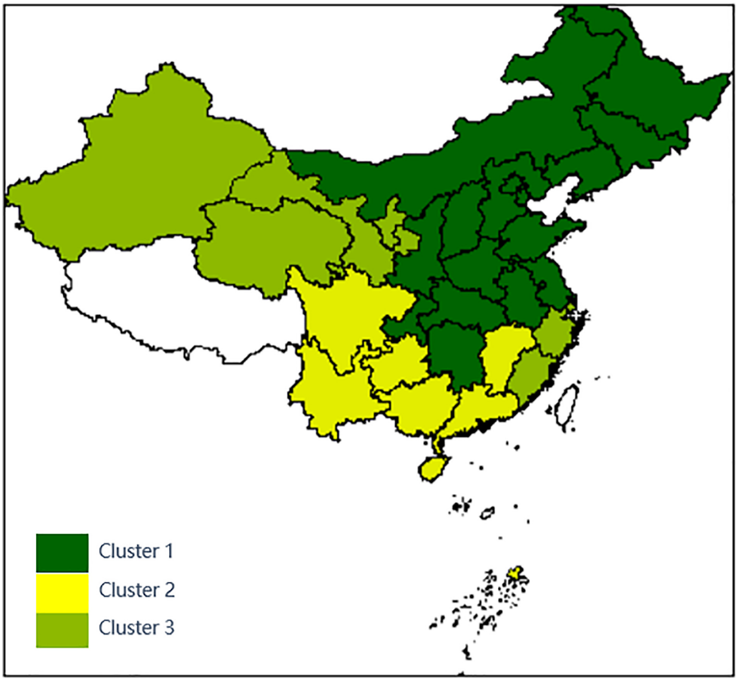


Figure 6.
Clustering results for TADPole algorithms

provinces, and use environmental regulations to stabilize the production. Though eggs production has a relatively high feed converting ratio, large poultry farms also have substantial local environmental impacts, such as manure discharge, water pollution, heavy metal pollution and air pollution (Cang *et al.*, 2004).

In 2016, Chinese government issued regulations to set up Livestock and Poultry Farming Prohibited Areas. However, the radical policy change caused high food price volatility in the aftermath (Figure 1). How to set up a flexible and sustainable regulation is a challenging issue for China.

7. Conclusions and policy implications

Price is an important signal for market equilibrium. Consumers, producers and governments make decisions according to the price (Sun *et al.*, 2017). Given the sheer geographic size of China, a lot of regional heterogeneities are observed, and policy-makers should take into account the regional heterogeneity and apply regionally targeted policies. Hence, how to cluster regional prices in China is an important research question. We introduce the DTW algorithm, a machine learning technique for clustering time series, to study regional price dynamics for eggs in China. Eggs are an inexpensive, nutritious and sustainable animal product. China is the largest country for eggs in the world both for production and consumption. Such a study has important policy significance.

We identified three clusters. The first cluster includes 12 provinces: Beijing, Tianjin, Hebei, Shanxi, Liaoning, Jilin, Heilongjiang, Jiangsu, Shandong, Henan, Chongqing and Shaanxi. The second cluster includes 2 provinces, namely, Anhui and Hubei. These two clusters have relatively high price volatility and are main egg production regions. The third cluster has relatively low price volatility, including Inner Mongolia, Shanghai, Zhejiang, Fujian, Jiangxi, Hunan, Guangdong, Guangxi, Hainan, Sichuan, Guizhou, Yunnan, Gansu, Qinghai, Ningxia and Xinjiang. These are main egg importing provinces.

We also ran a multinomial logit model to study the determinants of the clusters. We confirm that main production regions and some extended provinces belong to the first cluster and have relatively high price volatility. This is consistent with the literature and economic theory. Due to transaction costs, the importing areas may experience less price volatility (Meyer and von Cramon-Taubadel, 2004). In other words, if prices decline rapidly in exporting regions due to an oversupply in eggs, prices in importing regions will still remain relatively constant due to transaction costs barriers.

The clustering of Chinese provincial egg prices may promote an insightful understanding of price changes in the past few decades as well as improved comprehension of the determinants of fluctuation patterns at the provincial level. The work of our paper could facilitate policy-makers in developing target policy for a certain group of provinces. As a result, both government and the producers can be prepared for the possible changes, to reduce potential risk and ensure food security all the time. Particularly, the government should make policies to target the main production regions to stabilize their production, and environmental regulation could be a policy instrument.

Given the sheer size of the country and its large production and consumption market for eggs, China represents an influential economy, not only in the region but also globally, for its wide coverage of comprehensive and complicated situations. Consequently, the underlying novel way of egg price analysis might potentially be adapted for other countries, which are experiencing similar regional food price volatilities.

Further applications of the DTW algorithm might entail the pattern recognition of severe disease spreading. Countries confronting severe animal epidemics that spread rapidly between regions can resort to the DTW analysis to explore effective prevention measures based on common features derived from the more precise clustering results of DTW. Regional infection patterns might follow similar trends but start at diverse time points, indicating time shifts of dynamic time series.

Another area of application of the DTW algorithm might be in the analysis of drastic price volatilities, for instance, caused by a significant shock of production. In this scenario, diverse range of time shifts could exist and must be taken into account if different policy interventions are to be implemented in different clusters, while DTW could exactly identify the time shifts. These are issues of great concern that many countries are currently experiencing or urgently need to address.

We apply and demonstrate the DTW's superior performance in dealing with specific features of dynamic data by clustering prices of a vital agricultural product in China. Similar analysis and implementation of policy could be developed depending on the cases of each country. However, this also needs enough attention from the government and policy-making sectors.

References

Aghabozorgi, S., Shirkhorshidi, A.S. and Teh Ying, W. (2015), "Time-series clustering - a decade review", *Information Systems*, Vol. 53, pp. 16-38, doi: 10.1016/j.is.2015.04.007.

Alexander, P., Brown, C., Arneth, A., Finnigan, J. and Rounsevell, M.D.A. (2016), "Human appropriation of land for food: the role of diet", *Global Environmental Change-Human and Policy Dimensions*, Vol. 41, pp. 88-98, doi: 10.1016/j.gloenvcha.2016.09.005.

Bastianin, A., Conti, F. and Manera, M. (2016), "The impacts of oil price shocks on stock market volatility: evidence from the G7 countries", *Energy Policy*, Vol. 98, pp. 160-169, doi: 10.1016/j.enpol.2016.08.020.

Begum, N., Ulanova, L., Wang, J., Keogh, E. and Association for Computing Machinery (2015), "Accelerating dynamic time warping clustering with a novel admissible pruning strategy", *Paper presented at the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, University of Technology Sydney, Advanced Analytics Institute, Sydney.

Bellemare, M.F., Barrett, C.B. and Just, D.R. (2013), "The welfare impacts of commodity price volatility: evidence from rural Ethiopia", *American Journal of Agricultural Economics*, Vol. 95 No. 4, pp. 877-899, doi: 10.1093/ajae/aat018.

Berndt, D.J. (1994), "Using dynamic time warping to find patterns in time series", *KDD Workshop*, Vol. 10, pp. 359-370.

Cang, L., Wang, Y.J., Zhou, D.M. and Dong, Y.H. (2004), "Heavy metals pollution in poultry and livestock feeds and manures under intensive farming in Jiangsu Province, China", *Journal of Environmental Sciences*, Vol. 16 No. 3, pp. 371-374.

Chai, J., Wang, Y.R., Wang, S.Y. and Wang, Y.Y. (2019), "A decomposition-integration model with dynamic fuzzy reconstruction for crude oil price prediction and the implications for sustainable development", *Journal of Cleaner Production*, Vol. 229, pp. 775-786, doi: 10.1016/j.jclepro.2019.04.393.

Chen, J.M. and Rehman, M.U. (2021), "A pattern new in every moment: the temporal clustering of markets for crude oil, refined fuels, and other commodities", *Energies*, Vol. 14 No. 19, doi: 10.3390/en14196099.

Conradt, T., Gornott, C. and Wechsung, F. (2016), "Extending and improving regionalized winter wheat and silage maize yield regression models for Germany: enhancing the predictive skill by panel definition through cluster analysis", *Agricultural and Forest Meteorology*, Vol. 216, pp. 68-81, doi: 10.1016/j.agrformet.2015.10.003.

Dhini, A., Surjandari, I., Riefqi, M. and Puspasari, M.A. (2015), "Forecasting analysis of consumer goods demand using neural networks and ARIMA", *International Journal of Technology*, Vol. 6 No. 5, pp. 872-880, doi: 10.14716/ijtech.v6i5.1882.

Ding, I.J. and Shi, J.Y. (2017), "Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots", *Computers and Electrical Engineering*, Vol. 62, pp. 719-729, doi: 10.1016/j.compeleceng.2015.12.010.

Dmytrow, K., Landmesser, J. and Bieszk-Stolorz, B. (2021), "The connections between COVID-19 and the energy commodities prices: evidence through the dynamic time ing method", *Energies*, Vol. 14 No. 13, doi: 10.3390/en14134024.

Dong, H.J., Guo, X.M., Reichgelt, H. and Hu, R.Z. (2020), "Predictive power of ARIMA models in forecasting equity returns: a sliding window method", *Journal of Asset Management*, Vol. 21 No. 6, pp. 549-566, doi: 10.1057/s41260-020-00184-z.

Gardner, B.L. (1975), "Farm retail price spread in a competitive food-industry", *American Journal of Agricultural Economics*, Vol. 57 No. 3, pp. 399-409, doi: 10.2307/1238402.

Garnett, T. (2009), "Livestock-related greenhouse gas emissions: impacts and options for policy makers", *Environmental Science and Policy*, Vol. 12 No. 4, pp. 491-503, doi: 10.1016/j.envsci.2009.01.006.

Giorgino, T. (2009), "Computing and visualizing dynamic time warping alignments in R: the dtw package", *Journal of Statistical Software*, Vol. 31 No. 7, pp. 1-24, doi: 10.18637/jss.v031.i07.

Graskemper, V., Yu, X.H. and Feil, J.H. (2021), "Farmer typology and implications for policy design - an unsupervised machine learning approach", *Land Use Policy*, Vol. 103, doi: 10.1016/j.landusepol.2021.105328.

Graskemper, V., Yu, X. and Feil, J.-H. (2022), "Values of farmers – evidence from Germany", *Journal of Rural Studies*, Vol. 89, pp. 13-24, doi: 10.1016/j.jrurstud.2021.11.005.

Herrero, M., Havlik, P., Valin, H., Notenbaert, A., Rufino, M.C., Thornton, P.K., Bluemmel, M., Weiss, F., Grace, D. and Obersteiner, M. (2013), "Biomass use, production, feed efficiencies, and greenhouse gas emissions from global livestock systems", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 110 No. 52, pp. 20888-20893, doi: 10.1073/pnas.1308149110.

Jadhav, V., Reddy, B.V.C. and Gaddi, G.M. (2017), "Application of ARIMA model for forecasting agricultural prices", *Journal of Agricultural Science and Technology*, Vol. 19 No. 5, pp. 981-992.

Keogh, E. and Ratanamahatana, C.A. (2005), "Exact indexing of dynamic time warping", *Knowledge and Information Systems*, Vol. 7 No. 3, pp. 358-386, doi: 10.1007/s10115-004-0154-9.

Kraemer, M.U.G., Yang, C.H., Gutierrez, B., Wu, C.H., Klein, B., Pigott, D.M. ... and Open, C.-D.W.G. (2020), "The effect of human mobility and control measures on the COVID-19 epidemic in China", *Science*, Vol. 368 No. 6490, p. 493, doi:10.1126/science.abb4218.

Lemire, D. (2009), "Faster retrieval with a two-pass dynamic-time-warping lower bound", *Pattern Recognition*, Vol. 42 No. 9, pp. 2169-2180, doi: 10.1016/j.patcog.2008.11.030.

Maruejols, L., Höschle, L. and Yu, X. (2022), "Vietnam between economic growth and ethnic divergence: a LASSO examination of income-mediated energy consumption", *Energy Economics*, Vol. 114, doi: 10.1016/j.eneco.2022.106222.

Meila, M. (2003), "Comparing clusterings by the variation of information", in Scholkopf, B. and Warmuth, M.K. (Eds), *Learning Theory and Kernel Machines*, Springer-Verlag Berlin, Berlin, Vol. 2777, pp. 173-187.

Meyer, J. and von Cramon-Taubadel, S. (2004), "Asymmetric price transmission: a survey", *Journal of Agricultural Economics*, Vol. 55 No. 3, pp. 581-611, doi: 10.1111/j.1477-9552.2004.tb00116.x.

Nakagawa, K., Imamura, M. and Yoshida, K. (2019), "Stock price prediction using k-medoids clustering with indexing dynamic time warping", *Electronics and Communications in Japan*, Vol. 102 No. 2, pp. 3-8, doi: 10.1002/ecj.12140.

National Bureau of Statistics (2020), *Chiina Rural Statistical Yearbook*, China Statistics Press, available at: https://data.cnki.net/yearbook/Single/N2020120306.

Nijdam, D., Rood, T. and Westhoek, H. (2012), "The price of protein: review of land use and carbon footprints from life cycle assessments of animal food products and their substitutes", *Food Policy*, Vol. 37 No. 6, pp. 760-770, doi: 10.1016/j.foodpol.2012.08.002.

Pfisterer, F., Beggel, L., Sun, X., Scheipl, F. and Bischl, B. (2019), "Benchmarking time series classification - functional data vs machine learning approaches", *ArXiv*, abs/1911.07511.

Ratanamahatana, C.A. and Keogh, E.J. (2004), "Everything you know about dynamic time warping is wrong", *Third Workshop on Mining Temporal and Sequential Data*, Vol. 32.

Rehault-Godbert, S., Guyot, N. and Nys, Y. (2019), "The golden egg: nutritional value, bioactivities, and emerging benefits for human health", *Nutrients*, Vol. 11 No. 3, doi: 10.3390/nu11030684.

Sardá-Espinosa, A. (2019), "Time-series clustering in R using the dtwclust package", *The R Journal*, Vol. 11, p. 22, doi: 10.32614/RJ-2019-023.

Shahriar, S.A., Kayes, I., Hasan, K., Hasan, M., Islam, R., Awang, N.R., Hamzah, Z., Rak, A.E. and Salam, M.A. (2021), "Potential of ARIMA-ANN, ARIMA-SVM, DT and CatBoost for atmospheric PM2.5 forecasting in Bangladesh", *Atmosphere*, Vol. 12 No. 1, doi: 10.3390/atmos12010100.

Sun, F.F., Koemle, D.B.A. and Yu, X.H. (2017), "Air pollution and food prices: evidence from China", *Australian Journal of Agricultural and Resource Economics*, Vol. 61 No. 2, pp. 195-210, doi: 10.1111/1467-8489.12204.

Tian, X. and von Cramon-Taubadel, S. (2020), "Economic consequences of African swine fever", *Nature Food*, Vol. 1 No. 4, pp. 196-197, doi: 10.1038/s43016-020-0061-6.

Tian, X., Zhang, X.H., Zhou, Y.H. and Yu, X.H. (2016), "Regional income inequality in China revisited: a perspective from club convergence", *Economic Modelling*, Vol. 56, pp. 50-58, doi: 10.1016/j.econmod.2016.02.028.

Tsinaslanidis, P.E. (2018), "Subsequence dynamic time ing for charting: bullish and bearish class predictions for NYSE stocks", *Expert Systems with Applications*, Vol. 94, pp. 193-204, doi: 10.1016/j.eswa.2017.10.055.

Tweeten, L. (1979), *Foundations of Farm Policy*, The University of Nebraska Press, Nebraska.

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

Wang, H., Maruejols, L. and Yu, X. (2021a), “Predicting energy poverty with combinations of remote-sensing and socioeconomic survey data in India: evidence from machine learning”, *Energy Economics*, Vol. 102, doi: 10.1016/j.eneco.2021.105510.

Wang, H.J., Feil, J.H. and Yu, X.H. (2021b), “Disagreement on sunspots and soybeans futures price”, *Economic Modelling*, Vol. 95, pp. 385-393, doi: 10.1016/j.econmod.2020.03.005.

Wobbrock, J.O., Wilson, A.D. and Li, Y. (2007), “Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes”, *Paper presented at the 20th Annual ACM Symposium on User Interface Software and Technology*, Newport, RI.

Yin, Y. and Shang, P.J. (2014), “Modified multidimensional scaling approach to analyze financial markets”, *Chaos*, Vol. 24 No. 2, doi: 10.1063/1.4873523.

Yu, X.H., Liu, C., Wang, H.J. and Feil, J.H. (2020), “The impact of COVID-19 on food prices in China: evidence of four major food products from Beijing, Shandong and Hubei Provinces”, *China Agricultural Economic Review*, Vol. 12 No. 3, pp. 445-458, doi: 10.1108/caer-04-2020-0054.

Yu, X.H., Huang, Y.Y. and Wang, H.J. (2021), “Rethinking agricultural and rural development under the new pattern of domestic circulation”, *Journal of Huazhong Agricultural University (Social Sciences Edition)*, Vol. 3 No. 3, pp. 10-18 + 182-183.

Yu, X.H. (2014a), “Monetary easing policy and long-run food prices: evidence from China”, *Economic Modelling*, Vol. 40, pp. 175-183, doi: 10.1016/j.econmod.2014.03.029.

Yu, X.H. (2014b), “Raising food prices and welfare change: a simple calibration”, *Applied Economics Letters*, Vol. 21 No. 9, pp. 643-645, doi: 10.1080/13504851.2013.879281.

Zhou, D. and Koemle, D. (2015), “Price transmission in hog and feed markets of China”, *Journal of Integrative Agriculture*, Vol. 14 No. 6, pp. 1122-1129, doi: 10.1016/s2095-3119(14)60995-3.

Corresponding author
Xiaohua Yu can be contacted at: xyu@uni-goettingen.de