Abstract

Due to the complex and uncertain nature of financial market, investments are risky and the return becomes insecure. Consequently, smart hedging and trading strategies are essential for financial institutions to control the risk. The main aim of this research is to find exact trading process and to find appropriate algorithms in constructing portfolios for smart hedging when a company can not find one certain derivative that can hedge their potential loss. The scenario of this research is a risk management platform Stable who is issuing and underwriting derivative contracts over different commodities. Focusing on Stable's agricultural commodities, we consider a total of 344 agricultural products on the financial market and uses them as a pool and use futures as an instrument of hedging to propose two different strategies, which are trading with single futures using set benchmark to make decisions as well as trading with multiple futures by building a portfolio with highly related assets. It also gave advice on the amount and the timing for the company to do the trading in order to hedge their risk and to protect their profit.

Contents

1 Introduction **3**
- 1.1 Background . . . 3
- 1.2 Literature Review . . . 3
- 1.3 Motivation . . . 4
- 1.4 Overview . . . 5

2 Data Preprocessing **6**
- 2.1 Data Collection . . . 6
- 2.2 Normalisation by Logarithm Difference Index . . . 6

3 Inspiration from Current Trading Implementation **7**
- 3.1 Choices for Future Contracts . . . 7
- 3.2 Strategy Implementation . . . 7
- 3.3 Case Study . . . 9
- 3.4 Interpretation of Result . . . 10

4 Innovative Strategy **11**
- 4.1 Main Methodology with MLR Model . . . 11
  - 4.1.1 Traditional OLS Method & Case Study . . . 12
- 4.2 Selection of Basket Options . . . 14
  - 4.2.1 K-Means & Hierarchical Clustering . . . 14
  - 4.2.2 Interpretation of Clustering Result . . . 16
  - 4.2.3 Result Analysis of MLR Model based on clustering . . . 18
- 4.3 Trading Simulation . . . 19
  - 4.3.1 Explanation of Innovative Strategy . . . 19
  - 4.3.2 Transaction Amount & Timing . . . 20
- 4.4 Case Study . . . 23
- 4.5 Limitation . . . 24

5 Conclusion **24**

1 Introduction

1.1 Background

Over the past few years, volatile commodity prices and highly-risky exchange market were seen as significant global phenomena. The combination of supply in different commodities and inelastic demand means that unanticipated changes in one of both parts might generate large price swings, at least in the short term, which makes a reasonable hedging and trading process extremely important. Therefore, in the financial market, many firms are issuing different types of options or spreads every year, including Stable, in order to reduce the loss that might come along with those those unanticipated changes. Stable is currently issuing and underwriting more than 250 types of derivative contracts over 10 different commodities, which are virtually insurances to protect individual farmers or farms from huge loss of money when the price of their product decrease, with the form of spread options. Like a vanilla option, a spread option is a type of option that derives its value from the difference or spread between the prices of two or more asset [1]. However, the company itself also faces potential market risk exposures from these European or Asian style spreads in the Chicago Board of Trade (CBOT), Chicago Mercantile Exchange and Chicago Board of Trade (CME) and other Over the counter (OTC) markets. The company is currently in need of a trading strategy that can prevent these potential opportunities from happening to the current portfolios.

1.2 Literature Review

Many well-known strategies have been proposed before in order to find the best trading strategy for a certain entity, such as mean-variance strategy and momentum strategy. Markowitz proposed mean-variance strategy, which tried to optimize portfolio return over longer horizons [11]. On the contrary, Okunev and White mentioned that a short-term momentum effect can bring profitability in the foreign exchange market trading [13]. Both of these two strategies are about giving the trend of prices, in other words, rough prediction. Besides these two strategies, many also hold the view that buy-and-hold strategy is the most efficient strategy. Glassman and Hassett promoted buy-and-hold Strategy in their book, a passive and conservative strategy where investors buy and hold their stocks for a relatively long period, with no consideration of any fluctuation of price in the market [3]. Our finding is to give investors ideas about when to choose and how to choose basing on the direction they have already chosen, using mathematics calculation and formula derivation in order to find them.

Delta hedging utilizes hedging amount to precisely mimic the sensitivity of the derivative price with respect to Stable's asset. In delta hedging, delta has dynamic and oscillating variation with the asset price. This is the

reason why if the asset price have significant change in a very short time, the determined Δ will not be proper for hedging in the whole time period T. This property of Δ makes its re-balancing in time necessary in this strategy. However, it leads to much computational and operational cost. Meanwhile, delta hedge only protects against small price movements of the underlying but does not protect against larger movements of the underlying asset. When the underlying asset moves, the non-zero gamma will change the delta, resulting in re-hedging. Actually, an option's P&L depends on the P&L of delta, gamma, vega, etc. If merely considering delta hedge, only a small part of risk is hedged. Hence gamma hedge and other hedging strategies should also be thought over.

During the process of machine learning, Celebi and Aydin (2016) mentioned supervised learning especially with regression model is to add labels for training dataset and used for testing. Before development of machine learning, simple linear regression is frequently considered for forecasting. However, Choudhury et al. (2013) discovered that most stock data is complex and noisy in nature which interfere those regression methods to find significant results. Simply for unsupervised learning, distinctive clustering techniques are used to effectively carry out the quantitative analysis, screening out primary items from a basket of futures or options. It did not become a financial data-solving method until around two decades ago, which is able to rule out interferes from data sets without the need for labels (Li and Wong, 2019). According to Hansen et al. (2005), K-means method, whose standard algorithm is to find groups for the data with specific number for groups of all items represented by variable K, is the common strategy that companies choose to cluster data. It was first proposed by Stuart Lloyd in 1957. Choudhury et al. (2013) states that K-means clustering combined with neural network considers two values involving logarithmic return and daily underlying volatility in the case of multi-dimensional market data. Nevertheless, this method is restricted to adjust the suitable parameter K and adopt power initialisation, thus leading to high computational cost. It could be compared with hierarchical clustering, which is constructed in a tree structure for successive merging and hierarchical clustering is less frequently mentioned.

1.3 Motivation

As is known, those potential risk brought by these unanticipated changes can be hedged with derivatives such as options or futures. However, to take Stable as an example, it is relatively hard to find derivative of their product in the financial market. In order to solve this problem, we use highly related assets to build a portfolio in order to provide a strategy that can be used by risk management platform and give the specific operation which has been used. The main subject of this paper is derivatives of agricultural products,

nevertheless, the methods used in this paper can also be used for future research about other products like gold, heavy metal and energy source.

1.4 Overview

We proposed two trading strategies in this essay, with two very different train of thought basing on simulations of two different situations, yet with the same target to fully hedge the risk from spread options. In the first strategy, we use futures as an instrument to offset the risk that our issued spread options brought us, and use the spot price movement of our underlying asset to make our choice between short and long positions of the futures written on our underlying asset with good timing. In this strategy, with the idea of dynamic hedging, we decided to set a benchmark on the strike price which allow us to take different actions with our short or long positions, based on whether the spot price is above the benchmark or below the benchmark. We designed our algorithm with Visual Basic Application in order to realize this strategy.

We designed our second strategy under the situation that those options and futures written on our underlying assets are actually untradeable, in which case a portfolio of closely related assets that can be traded is in need to offset the potential loss our spread might bring us.In the process of finding such optimal hedging portfolio, the essential indicator used is log Return which is directly calculated from the data of the history monthly close price of the products in the pool. Following this, K-means and Hierarchical clustering are used to select the relatively higher correlated ones with Stable's commodities from the pool. The Silhouette Coefficient is used to compare between these two models. The results show that hierarchical performs better as expected. Finally, based on the clustering result of Hierarchical, the optimal hedging portfolio is constructed using two linear regression models whose results consist. Then we modeled a linear relationship between a scale dependent variable and independent variables. Specifically, basing on the linear behavior of data, correlated dependent ones can be predicted by the multiple linear regression. After finding our portfolio with such algorithm using Python as our platform, we again use futures to reduce our risk. This time, instead of single ones, however, we use our portfolio as a base to choose futures from every single one of our chosen assets and build a new portfolio of futures. We then realize the reduction of our risk by trading this portfolio of futures on their expiration data.

2 Data Preprocessing

2.1 Data Collection

The two main data sources of this research are Stable and Thomson Reuters Eikon (TRE). Sikacz and Wolczek (2018) mentioned that TRE is a database that provides the latest and accurate financial information in over 400 stock exchange markets [14].

All the data from these two sources are time series which is defined as a series of data that are recorded in equal time intervals according to Bhanja and Das [7]. In detail, the initial data from Stable includes the average close prices per month of ten agricultural commodities. The history price of these commodities available in TRE is from July 2009 to March 2019 and this becomes the time period that is considered. Following this, valid data is obtained through spline interpolation to complete empty place.

2.2 Normalisation by Logarithm Difference Index

Time series data are inherently nonlinear and fluctuate irregularly. Thus, before importing data into clustering algorithms, normalisation is very necessary [7]. Log return can reduce the scale of variable into unified smaller range. According to Taylor expansion about $x = 0$, $f(x) = \sum_{j=0}^{\infty} f^{(j)}(0) \times \frac{x^j}{j!}$, the formula about return $r_{t+1} = \ln \frac{P_{t+1}}{P_t} \approx \frac{P_{t+1}-P_t}{P_t}$ is deduced and logarithm difference index represents monthly return rate from financial perspective.

Each return's trend is consistent with normal distribution, which is the condition for obtaining correlation and returning back to original price. Derivation from fitting return well to relationship of price is presented.

$$\begin{aligned} r_{t(A)} &\approx \theta_1 r_{t(1)} + \theta_2 r_{t(2)} + ...\theta_n r_{t(n)} \\ r_{t(A)} &= \ln \frac{y_{t(A)}}{y_{t-1(A)}} \ , \ r_{t(i)} = \ln \frac{x_{t(i)}}{x_{t-1(i)}} \end{aligned} \tag{1}$$

$$\prod_{i=1}^{n} (x_{t(i)})^{\theta_i} \approx \mathrm{e}^{r_{t(A)}} \cdot \prod_{i=1}^{n} (x_{t-1(i)})^{\theta_i} \approx \frac{y_{t(A)}}{y_{t-1(A)}} \cdot \prod_{i=1}^{n} (x_{t-1(i)})^{\theta_i} \tag{2}$$

$$y_{t(A)} \approx \frac{y_{t-1(A)} \prod_{i=1}^{n} (x_{t(i)})^{\theta_i}}{\prod_{i=1}^{n} (x_{t-1(i)})^{\theta_i}} \tag{3}$$

Here, $x_{t-1(i)}$ is known historical data. Based on the fact that price of underlying asset follows random walk causing unstationary process and log return's sequence is stationary, log return is reliable for linear fitting.

Log return is not only used in clustering, but also applied in multivariate linear regression as well. Heteroscedasticity and collinearity exist in linear regression model and log return could solve the issue. Then, data set is used to construct hedging portfolios in this research.

3 Inspiration from Current Trading Implementation

3.1 Choices for Future Contracts

The idea of this strategy is hedging the potential loss of a short bear spread by shorting and longing futures. The commodity which the bear spread is based on should at least satisfies one of the following conditions:

i There are tradable futures corresponding to this commodity.

ii If there is no corresponding tradable future, we can find substitute futures in the same category with similar price movement.

The assumptions above ensure that future prices and commodity prices have high correlation so that price movement of the commodity is utilized as signals to trade futures.

3.2 Strategy Implementation

Given appropriate futures, designing trading strategy is conducted. Before that, it is important to demonstrate why futures can be used to hedge spread options. Figure 1 is about short position of bear spread with strike $K_1 = 60$ and $K_2 = 120$, as well as a future contract with settlement price 120.

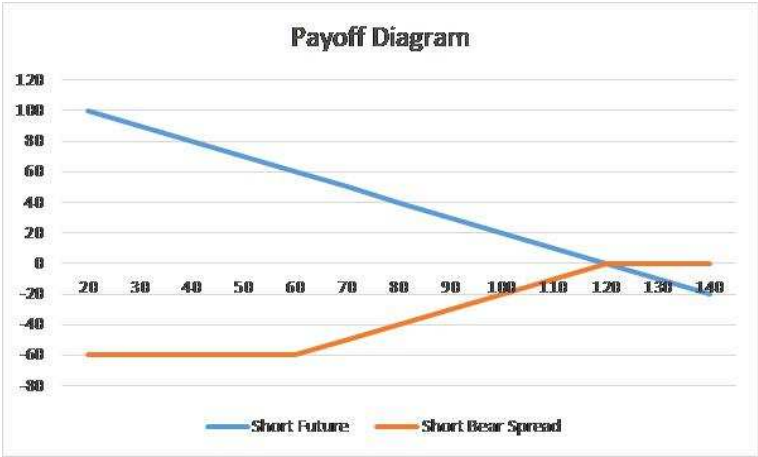

Figure 1: Payoff of the Short Bear Spread & short future

There will be a loss when the end price is below 120, so some derivatives should be added to compensate that loss. One possible candidate is shorting a future contract.

Clearly, opposite trend below 120 indicates that shorting a future are capable of perfectly hedging the risk of bear spreads issued. However, by

entering the short position of futures, the portfolio has been exposed to a new risk. If the end price of the asset actually rises higher than the settlement price of the future contract, there will be a loss. This is unacceptable since the ideal hedging strategy should be market neutral, which can profit from both increasing and decreasing prices in one or more markets[4].

The direct approach to this is entering a long position of futures in order to offset the previous short position if the future price goes up. Rise and fall of price can be quantified by a benchmark, which is set to be the higher strike price of spreads in this strategy. When the commodity price drops below the benchmark, the spread option starts losing money, which gives a signal to short futures so that we can compensate losses of the spread. After that if the commodity price moves above the benchmark, a signal to long is given in order to offset the current short position. Later, if the spot price drops below the benchmark again, signals to enter short position will be released again.

The long and short positions of futures are decided by the commodity price movement, which is based on the idea of dynamic hedging [8]. For every month, Algorithm 1 illustrates how the strategy works, check whether the commodity price crosses the benchmark, and then decide the operation.

Algorithm 1 Strategy Pseudocode

IF $P_1 < K_2$ **THEN**
  Short Future with F_1 at t_1
ELSE Do not change current position at t_1

REPEAT
IF $P_i - 1 < K_2$ and $P_i > K_2$
  THEN Long Future with F_i
  ELSEIF $P_i - 1 > K_2$ and $P_i < K_2$
    THEN Short Future with F_i
  ELSE Do not change current position at t_1
$i \leftarrow i + 1$
UNTIL i=duration

After implementing the trading strategy, two specific cases and corresponding outcome are provided, one of which is winning and the other is losing. The following diagram shows the commodity price and future price movement.

Table 2: Breakdown Results for different Maturities

Maturity	3	4	5	6	7	8	9	10	11	12
Spread Payoff	-2591.38	-3166.89	-3915.058	-4452.69	-4932.79	-6157.13	-5384.86	-5417.83	-5375.49	-5954.00
Strategy Payoff	-1879.48	-2110.72	-2848.27	-3375.49	-4039.99	-5112.91	-4820.51	-5313.24	-5 550.94	-6374.89
Payoff Difference	711.9	1056.17	1066.79	1077.2	892.8	1044.22	564.35	104.59	-175.45	-420.89
Payoff Difference Factor	0.27	0.33	0.27	0.24	0.18	0.17	0.10	0.02	-0.03	-0.07
Winning ratio	0.58	0.54	0.53	0.55	0.54	0.52	0.49	0.46	0.45	0.45

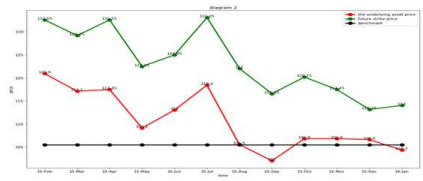

Figure 2: Winning Example

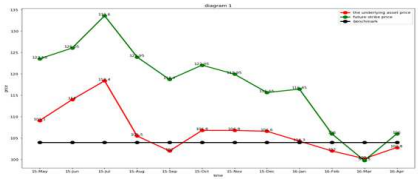

Figure 3: Losing Example

3.3 Case Study

The performance of our strategy using historical data should be tested. The data we used are all European bear put spread options issued by the company from August 2010 to May 2019, 15127 spreads in total. The strategy are implemented into all those spreads and payoff is calculated for each spread.

The performance of the strategy is evaluated by the Total Payoff Difference, Payoff Difference factor and Winning Ratio. Payoff difference indicates how much loss has the strategy effectively hedged and payoff difference factor relates the amount of payoff per unit of risk, with values close to one indicating an effective hedging system in Table 1.

Table 1: Overall Back Test Results

Total payoff of spreads	-47348.12489
Total Payoff of the strategy	-41426.44489
Total Payoff Difference	5921.68
Payoff Difference Factor	0.12507
Overall Winning Ratio	50.2145%

After implementing the strategy, the total loss has been reduced by nearly 6000 (12.507 % of the total loss). Among all 15127 spreads issued, around 50% have reduced loss after implemented the strategy. And after presenting the overall performance, let us break down the results according to different maturity dates, ranging from 3 months to 12 months in Table 2.

Normalized results (Payoff Difference Factor and Winning Ratio) are more useful when comparing performance discrepancy among different maturities. Figure 4 and 5 illustrate the Payoff Difference Factor and the Winning Ratio for all maturities.

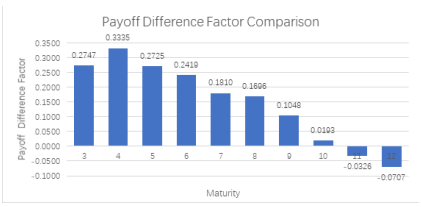

Figure 4: Payoff Difference Factor Comparison

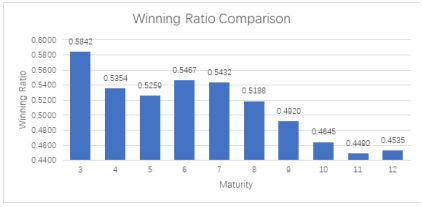

Figure 5: Winning Ratio Comparison

For both metrics, there are downward trend as maturity time increases. This indicates the strategy works more effectively on short-term spreads than long-term spreads.

3.4 Interpretation of Result

Although the strategy has successfully hedged a proportion of losses less than 13%, its performance is far below our expectation. An ideal strategy should have the capacity to eliminate approximately all the potential losses regardless of the price movements, in other words, achieving market neutral. There are two reasons accounts for the incompetency of our strategy. The first comes from the design of the strategy and the second is due to the input data.

In our strategy, signals are given only if the price has already dropped below the benchmark. At that moment, there is already a loss being generated by difference between benchmark value and current price. What the strategy can do is to stop further losses if the price continues decreasing but it can never prevent losses from happening. This can be a fatal flaw in some cases, if the price drops significantly from one observation point above the benchmark to the adjacent one below the benchmark, the initial loss can be extremely large.

The data we used in back testing are also likely to contribute to large initial losses. The interval between adjacent observation points is one month, which is excessively long. The price can change dramatically in one month's time especially for agriculture commodities, whose prices have strong seasonal fluctuations[12]. Therefore, the drastic drops or rises between two adjacent observation points are more likely to happen, which will possibly lead to high initial losses. This can also explain why short-term spreads can be hedged by the strategy more effectively than long-term ones. For long-term contracts, they have experienced more price movements before expire

and every such movement may lead to initial losses.

4 Innovative Strategy

Many drawbacks of the strategy for trading single future are discussed before. One drawback is that corresponding tradable future cannot be found. Furthermore, even if there are tradable corresponding futures, the previous model is highly dependent on the similarity of the future and underlying assets. limitations of the model generate the necessary motivation. Therefore, we advance an idea that the future could be replaced by different kinds of future combination.

Under price relationship, we filtrate a larger amounts of futures contracts for diverse commodities in the same category around the world. Hong J. said that machine learning algorithms such as OLS and Lass methods showed superior predictive ratios for multi-variable predictive risk [5]. As a result, the shadow index, can be simulated well with best combination from all possible portfolios involving highly-related products to targeted underlying assets and used to work on the hedging standard.

4.1 Main Methodology with MLR Model

Linear regression algorithm that can find the future combination is introduced here. It is used to model a linear relationship between a scale of dependent variable and independent variables. Most quantitative analysts build their alphas strategies and risk models by the multi-factor regression theory. Especially, the multiple linear regression can predict correlated dependent ones based on the linear behaviour of data.

For a data set $\{y_i, x_{i1}, ..., x_{ik}\}_{i=1}^{n}$ of n time intervals, the model assumes that it is evident to have linear relationship between y_i and p-vector of x. Thus, OLS model with k variables takes the form that:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i, \ i = 1, 2, ..., n \tag{4}$$

Matrix form is $y = X\beta + \epsilon$. It is clear to gain the coefficients $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_i)$ based on n-dimension data $X = (X_1^T, X_2^T, ..., X_i^T)$ and $y = (y_1, y_2, ..., y_i)$.

In the MLR model, we take the same category underlying assets since the relationship is excepted to be strongly linear. Although the needed futures are not easily accessible, models constructed by multiple factors show higher stability and robustness. For the same reason, high-dimensional futures combination significantly improve the fit of the regression model, so that single linear regression is abandoned.

4.1.1 Traditional OLS Method & Case Study

OLS is a popular regression method to estimate unknown coefficient in statistics. For Equation 5, we are supposed to find the best result rather than get the exact solution. In the sense of solving the quadratic minimization problem,

$$\hat{\beta} = \arg\min_{\beta} S(\beta) \tag{5}$$

As well, R-Square and Root mean square error (RMSE) are regarded as the judgement standard of our whole model. The R square tests fit result of the regression model, and the RMSE indicates the accuracy of the prediction model for test dataset.

After data processing, program traverses C_{127}^{n} (525500 times loop) n-variate combinations of 127 variables selected from 344 products to reduce the number for loop and selects the best portfolio which the R^2 is largest. The algorithm performs linear regression on each combination, so each loop regression computation avoids the error of OLS processing on big datasets.

Moreover, the simulated result is analysed and visualized. Firstly, we find some contradiction to balance fitting degree and prediction accuracy. The R^2 indicates fitting degree and RMSE indicates testing accuracy. From experimental data, Figure 6 is obtained.

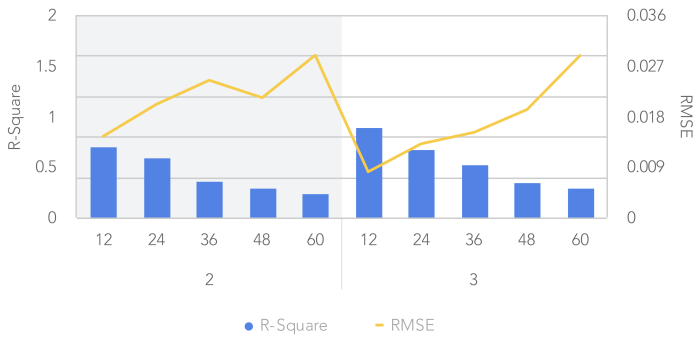

Figure 6: Statistics Balance between R-Square and RMSE

Table 3: Indicators influence on R-Square and RMSE

Indicator	R^2	RMSE
Training Period	-	-
number of variables	+	+

Additionally, as training period increase or the number of variables decrease, the fitting degree will be worse but the testing accuracy will be better, and vice versa. In this case, both R^2 and the RMSE will smaller. In particular, the abnormal RMSE is considered as signal of overfitting. We conclude in Table 3, where "+" means uptrend and "−" means downtrend.

After backtesting, we find the situation that keeps the balance of R^2 and RMSE: 4 years and 3 variables. Furthermore, OSR is regarded as target variable and linear regression is conducted. The best one is made up by Spices Herb NTMc2', 'Corn Cc2' and 'Rapeseed C0Mc1'. As well fitting details (Table 4) and graph (Figure 7) are following,

Table 4: Details of Optimal Fitting Result

Assets	Spices Herb NTMc2	Corn Cc2	Rapeseed C0Mc1
Coefficients	-0.123	0.237	0.430
R^2	0.23869		
RMSE	0.16736		

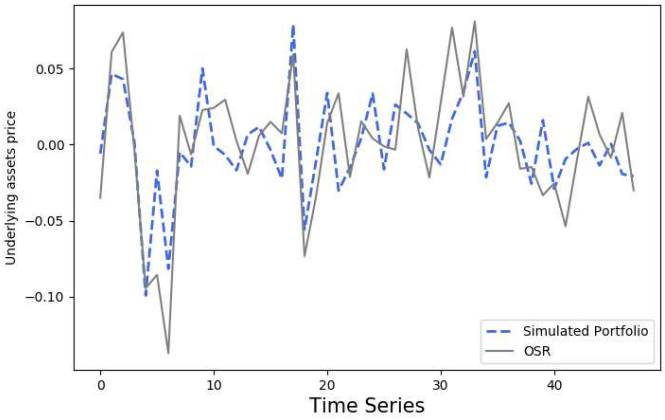

Figure 7: Graph of Optimal Fitting Result

Then, the table is presented with significant testing with t-test and F-test. Additionally, * means .05 level, ** means .01 level and *** means 0.001 significance level.

	Estimate	Std. Error	t-value	$Pr(>\|t\|)$		
Intercept	0.001737	0.001440	0.408	0.006847	**	
NTMc2	-0.122571	0.126071	2.186	0.032159	*	Multiple
Cc2	0.237334	0.193006	-3.9113	0.000209	***	
Repeseed	0.429659	0.117991	11.888	$< 2e^{-16}$	***	

R-Square: 0.54793, Adjusted R-Square: 0.52962

F-statistic: 1956 on 4 and 221 DF, p-value: $< 2.2e^{-16}$

In order to obtain smooth time-series data, we usually find the log return. For t-test, the p value of intercept and all coefficients are smaller than 0.05, so we can reject H_0 at 1% level in favour of $H_1 : \beta_1 > 0$.

- H_0: x_i has no relationship with y when [other x variables] are present in the model.
- H_1: x_i has a significant relationship with y when [other x variables] are present in the model.

Thus, there exists a very strong evidence between target price and each future indicator. For F-test, the results are $p - value < 2.2e^{-16} < 0.05$. So the regression relationship does usefully explain some of the variation in data. Above all, the model is approved.

4.2 Selection of Basket Options

4.2.1 K-Means & Hierarchical Clustering

Instead of iterating whole C_{127}^n (525500 times loops), clustering is utilized based on effective time complexity. Additionally, clustering technique is intelligent to determine time and total number of futures simultaneously without the necessity to balance R-square and RMSE. In addition, both OLS and gradient descent method could be used to find parameters for each product constituting portfolio after different clustering algorithms involving K-means & hierarchical clustering. Clustering analysis selects high-related products, reducing the amount of data for MLR model.

Table 5: Comparison between K-means & Hierarchical Clustering

	K-means	**Hierarchical(single linkage)**
Distance	Only limited to Euclidean Distance [15]	Apply to all distance standard such as correlation distance [16]
Procedure	1. Random initialization for K points 2. Assign remaining points to the nearest cluster 3. Update centroid of K clusters using arithmetic mean of each dimension of vector $\overrightarrow{\mu_j} = \frac{1}{n_j} \cdot \sum_{x_i \in j} \overrightarrow{x_i}$ 4. Termination Condition: $\left\|\vec{\mu} - \vec{\mu'}\right\| < 10^{-10}$ And condition is satisfied: $\|\overrightarrow{\mu_{k+1}} - \overrightarrow{\mu_k}\| < 10^{-10}$	1. each data point is initialized as an individual cluster and distance matrix whose entries are distance between two arbitrary clusters is computed; 2. Sort pairwise distance from minimum to maximum; 3. merge two closest clusters into one cluster and then distance matrix is updated by linkage function; 4. Repeat merging process until finally all the points fall into the same cluster.
Time Complexity	$O(nkdi)$, where i is number of iterations, n denotes number of data points and K is number of clusters, d represents the fixed dimension.	$O(n^2 \log_2 n)$, given n data points, and collection of pairwise distances [17]
Predetermine k value	Elbow method	
Visualization	PCA 3-D figure	Dendrogram

Elbow method to determine the optimal K value in K-means clustering algorithm

In K-means clustering method, the value of K should be fixed by Elbow method. Using elbow method in the following plot, the point with the largest variation for slope is called elbow point and its horizontal coordinate means the optimal value for K. Figure 8 shows evidence that $K = 5$ is an appropriate choice for dataset including deadweight cattle.

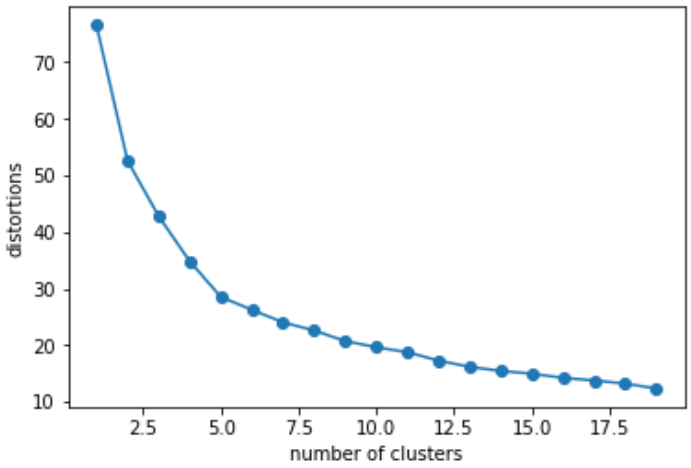

Figure 8: Line Chart of Elbow Method for Deadweight Cattle

4.2.2 Interpretation of Clustering Result

Agglomerative hierarchical clustering is a bottom-up mechanism conducted by merging small clusters into large one and single linkage hierarchical clustering is commonly used [2]. In Figure 9, dendrogram visualizes the hierarchical clustering process in a bottom-up mechanism of binary tree. At the bottom, there are 344 commodities and ten underlying assets. The bottom-up path shows the order for merging.

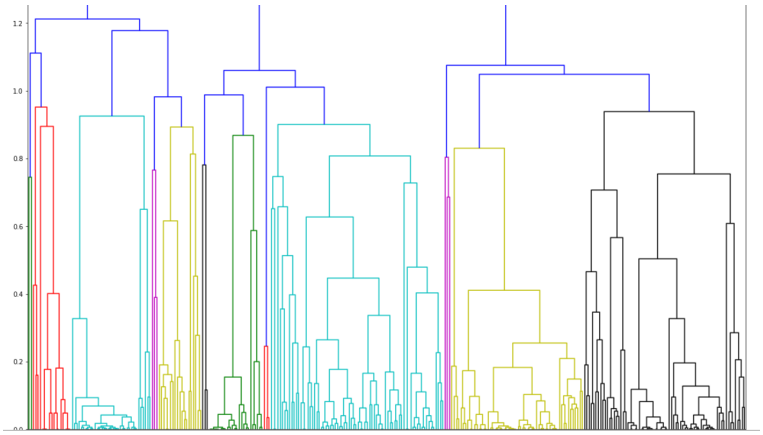

Figure 9: Commodity Sorting by Hierarchical Clustering (Single Linkage)

If performing a horizontal cut, the branches intersecting with cut rep-

resent different clusters. Once the distance reaches a critical standard, the combination will occur. Thus hierarchical clustering result is in Figure 10.

Objective Asset	Feed Wheat	Feed Barley	Milling Wheat	OSR	Milk	Deadweight Cattle	Lamb Deadweight	Pig	AN Fertiliser	Red Diesel
	corr=0.7	corr=0.7	corr=0.7	corr=0.7	isolated	corr=0.55	corr=0.7	isolated	corr=0.6	corr=0.6
	WheatBL2c1	WheatBL2c1	CornCc1	RapeseedCOMc1		Orange Juice0Jc10	Spices HarbNTMc2		RubberSRUc2	Timber1LBc6
	WheatBL2c2	WheatBL2c2	CornEMAc1	CornCc1		Orange Juice0Jc11	CornCc2		RubberSRUc3	RubberSTFc2
	WheatBL2c3	WheatBL2c3	CornEMACc1	RapeseedCOMc2		Orange Juice0Jc12	RapeseedCOMc1		CornCc1	RubberSTFc10
	WheatLWBc1	WheatLWBc1	WheatBL2c2	Soybean Meal 1SMcS		Orange Juice0Jc13			CornCc2	RubberSTFc11
Highly related asset from hierarchical			WheatBL2c3	RapeseedRSC3		Orange Juice0Jc14			DeadwheatLCc1	RubberSTFc12
			WheatKWc2			RiceRRc1				RubberSTFc7
						RiceRRc2				RubberSTFc8
						RiceRRc3				RubberSTFc1
						RobustaLRCc4				
						RobustaLRCc6				
						SugerLSUc8				
						Spices HarbWOMc4				

Figure 10: Hierarchical clustering result based on Insurance's commodity

However, for K-means clustering, each time, one of commodities is individually put into dataset and form 345 samples for later clustering. Similar result table will be obtained. After k-means clustering, observations are conducted projection into 3 orthogonal components using Principle component analysis (PCA), which is visualized in 3D space from Figure 11.

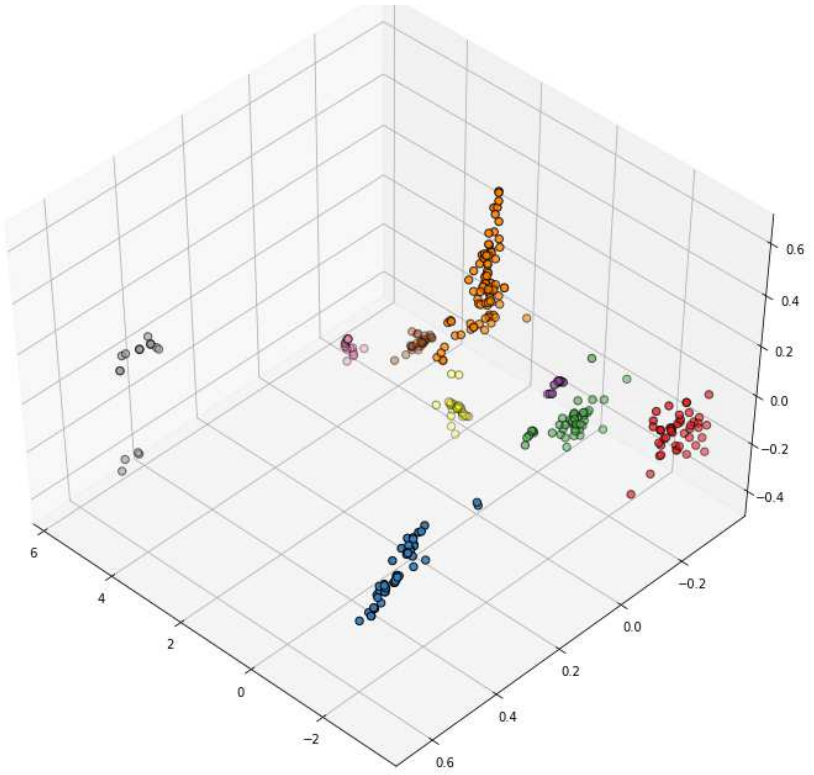

Figure 11: 3D K-means result by PCA

Silhouette Coefficient method for assessment

Silhouette Coefficient, is an evaluation method for measuring clustering effect [6]. According to the average intra-cluster dissimilarity $a(i)$ and average inter-cluster dissimilarity $b(i)$ of sample i, $s(i)$ is called the Silhouette coefficient of clustering. Based on research from Lleti et al. [10], the Silhouette coefficient is expressed as:

$$s(i) = \frac{b(i) - a(i)}{max\left\{a(i), b(i)\right\}} \tag{6}$$

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, a(i) < b(i), \\ 0, a(i) = b(i), \\ \frac{b(i)}{a(i)} - 1, a(i) > b(i) \end{cases} \tag{7}$$

If $s(i)$ is close to 1, then the clustering of commodity i is justifiable, and vice versa (Aranganayagi and Thangavel, 2007). $s(i)$ is applied in analyzing the whole effect for all clusters. The silhouette coefficient of both methods are presented below.

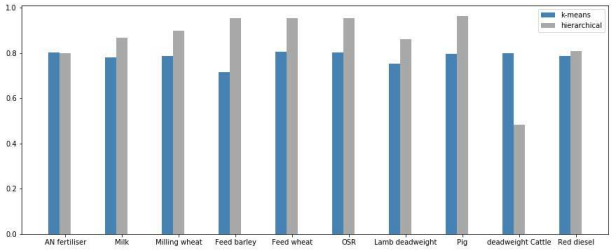

Figure 12: Silhouette Coefficient method for comparison about algorithms

As is shown in Figure 12, in spite of hierarchical clustering for deadweight cattle with relatively low value, the effect of hierarchical clustering generally outweighs K-means clustering based on effective positive Silhouette Coefficient.

4.2.3 Result Analysis of MLR Model based on clustering

Consequently, when selecting highly-correlated products for hedging, the result of Hierarchical clustering is adopted. Taking OSR as an example, 5 samples are finally in the same cluster with OSR which are 'Rapeseed COMc1', 'Rapeseed COMc2', 'Corn Cc1', 'Soybean Meal 1SMc8' and 'Rapeseed RSC3'. After multivariate linear regression model, parameters $\theta_1, ..., \theta_5$ is obtained to make up portfolio.

$$[\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6] = [0.003, 0.377, -0.037, -0.014, -0.051, 0.158] \quad (8)$$

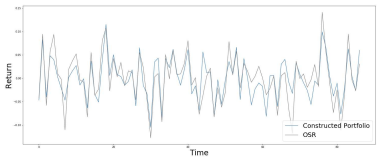

Figure 13: Training Set for 8 Years

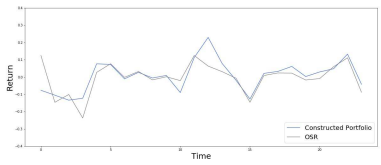

Figure 14: Testing Set for 2 Years

From OLS result, the fitting graph of train set (Figure 13) and test set (Figure 14) are shown below. In the figure 14, two lines fit well which can show the accuracy of the coefficients. Overall, return rate of this constructed portfolio has a similar trend as the target commodity OSR and can be used for hedging. In particular, to measure how well a hedging portfolio matches the movement of the targeted commodity in finance, R squared can be used [9]. Consequently, $R^2 = 0.690$ indicates a positive feedback of the regression.

4.3 Trading Simulation

4.3.1 Explanation of Innovative Strategy

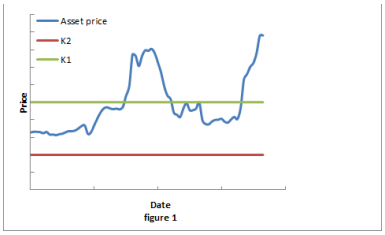

Figure 15: Price of the Underlying Asset

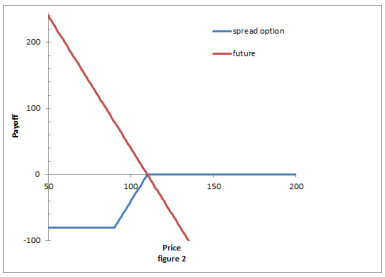

Figure 16: Payoff Diagram of the Spread Option and Future

After constructing the appropriate futures combinations, this section will introduce the strategy that trade the combination to hedge risks. It is known that agricultural insurance company have several spread options and we only consider European bear put spread option in our model. From Figure 15, the loss of the spread option could be expressed by the distance of the underlying asset price and the higher strike price K_1. The payoff could be directly seen in Figure 16. The non-constant part would be the only part we considered as illustrated later in the assumption. The red line

represents the portfolio profit, which can cover the loss of the spread option. Additionally, the future price in order to hedge the loss need to be find using the strategy mentions later. After that, things require to confirm is the quantity and time that we hold the portfolio and the individual future amount to buy or sell in it.

4.3.2 Transaction Amount & Timing

Optimal Amount

This part is about the specific trading strategies based on the assumptions as explained above. It is assumed that there is a strong similarity with the return rate of the underlying asset (A) and the portfolio (P) created in the model, which is shown in Equation 9 and r_t^A and r_t^P represent the return rate of A and P respectively.

$$\begin{aligned} r_t^A = r_t^P \ &\rightarrow e^{r_t^A} = e^{r_t^P} \\ \frac{x_t^A}{x_t^P} &= \frac{x_{t-1}^A}{x_{t-1}^P} \end{aligned} \tag{9}$$

where x_t^A and x_t^P are the price of A and P at time t correspondingly. Then Equation 10 is attained.

$$x_t^A = \frac{x_{t-1}^A}{x_{t-1}^P} x_t^P \tag{10}$$

There is a relationship between the price of A and P at time t, which informs the relative quantity of portfolio to be hold corresponds to the price of A and P. The exact amount of each future required has been indicated by the coefficient of the multiple linear regression equation.

Optimal Timing

From Equation 11,

$$\frac{x_t^A}{x_t^P} = \frac{x_{t-1}^A}{x_{t-1}^P} = \frac{x_{t-2}^A}{x_{t-2}^P} ... = \frac{x_{t-n}^A}{x_{t-n}^P} = M \tag{11}$$

$\frac{x_{t-1}^A}{x_{t-1}^P}$ is a constant number M based on the assumption, but it is obvious that the R-square from the regression model is not satisfied the assumption perfectly. To minimize the error, average value of M, called $\bar{M}$ is used. Combined with Equation 12, the equation is obtained.

$$x_t^A = \bar{M} \cdot x_t^P \tag{12}$$

α is set as the time to buy the portfolio. Then the profit g from portfolio is

$$g = x_\alpha^P - x_t^P \tag{13}$$

In order to hedge the loss, the total gain (G) need to be equivalent to the loss (L). From Equation 12 and 13, we can get

$$G = g \cdot \bar{M} = (x_\alpha^P - x_t^P) \cdot \bar{M} \tag{14}$$

With K_1 (in Figure 15) being the higher strike price of the spread option, we also know from our assumption that the loss L we need to hedge with our profit is

$$L = K_1 - x_t^A \tag{15}$$

Because we need, at least, our profit shall cover the loss, which means $G = L$.

$$\begin{aligned} x_\alpha^P \cdot \bar{M} - x_t^A &= K_1 - x_t^A \\ x_\alpha^P &= \frac{K_1}{\bar{M}} \end{aligned} \tag{16}$$

Because constant $\bar{M}$ could be calculated and K_1 is known, also with the coefficients gains from the regression equation we can therefore get the individual future prices in our portfolio. By treating these future prices as the benchmark, we decide the time α we need to buy the portfolio.

Practical Application for Each Components

After the relative quantity and time to enter a short or long position of future portfolio have been established, the individual relative quantities of individual future would be derived in this section. From the multiple linear regression with portfolio involving 3 kinds of futures in total, Equation 17 is as follows.

$$r_t^A = a_1 r_t^1 + a_2 r_t^2 + ... + a_i r_t^i \ \rightarrow r_t^A = a_1 r_t^1 + a_2 r_t^2 + a_3 r_t^3 \tag{17}$$

where a_1, a_2, a_3 are the coefficient of individual future price, which could be calculated by the algorithm. Log return rate(r) is defined in the Equation 18.

$$\ln \frac{x_t^A}{x_{t-1}^A} = a_1 \ln \frac{x_t^1}{x_{t-1}^1} + a_2 \ln \frac{x_t^2}{x_{t-1}^2} + a_3 \ln \frac{x_t^3}{x_{t-1}^3} \tag{18}$$

Where $x - 1$ represents the transaction one period earlier. After a simple operation, we can get Equation 19.

$$\frac{x_t^A}{x_{t-1}^A} = \frac{(x_t^1)^{a_1} \cdot (x_t^2)^{a_2} \cdot (x_t^3)^{a_3}}{(x_{t-1}^1)^{a_1} \cdot (x_{t-2}^2)^{a_2} \cdot (x_{t-1}^3)^{a_3}} = \frac{x_t^P}{x_{t-1}^P} \tag{19}$$

Then the denominator and numerator of the fraction would be showed in Equation 20.

$$\begin{aligned} x_{t-1}^{P} &= n \cdot (x_{t-1}^{1})^{a_1} \cdot (x_{t-1}^{1})^{a_2} \cdot (x_{t-1}^{3})^{a_3} \\ x_{t}^{A} &= n \cdot (x_{t}^{1})^{a_1} \cdot (x_{t}^{2})^{a_2} \cdot (x_{t}^{3})^{a_3} \end{aligned} \tag{20}$$

Where $x_{t-1}^{1}, x_{t-1}^{2}, x_{t-1}^{3}$ could be find in the database from insurance company. Since $x_{\alpha}^{P} = \frac{K_1}{M}$, with known K_1 we can get the result of x_{α}^{P}. Additionally, the sign of the coefficient will only show the opposite position we enter of related futures. Due to Equation 20 and the corresponding algorithm, the time α can be obtained finally, which we assume n=1 and we could use α express t, shown in Equation 21.

$$x_{\alpha}^{P} = (x_{\alpha}^{1})^{a_1} \cdot (x_{\alpha}^{1})^{a_2} \cdot (x_{\alpha}^{3})^{a_3} \tag{21}$$

And the total amount of each future can be expressed by

$$M_i = \frac{a_i}{a_1 + a_2 + a_3} \cdot \bar{M} \tag{22}$$

Figure 17 shows the main idea of our strategy and includes the parameters we have used.

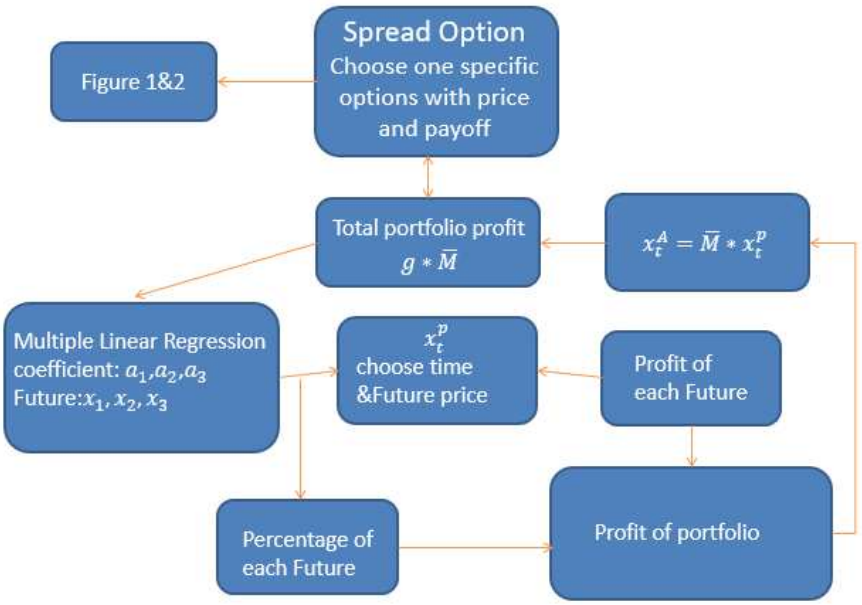

Figure 17: Main Calculations

4.4 Case Study

In order to better understand these equations, an example is as follows. From the results of clustering and multiple linear regression model, the best three assets that fit Lamb Deadweight are Rubber, Soybean and Wheat in Table 6, with corresponding coefficients 1.76285968, -0.64248426 and -0.99819122.

Table 6: Detail of Case Study

Assets	Spices Herb NTMc2	Corn Cc2	Rapeseed C0Mc1
Coefficients	-0.123	0.237	0.430
R^2	0.23869		
RMSE	0.16736		

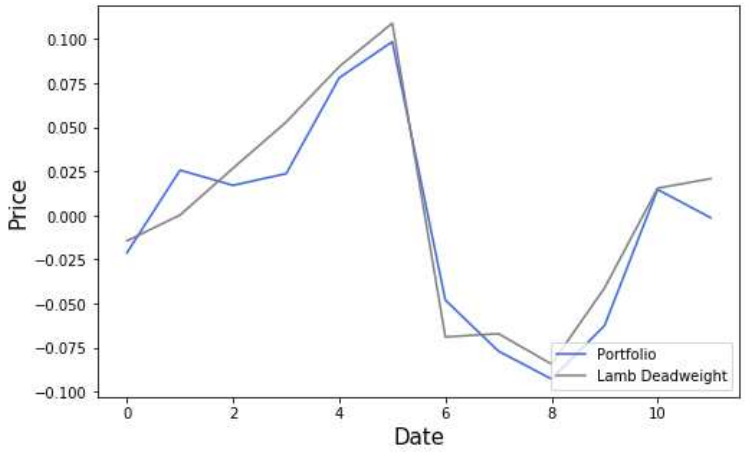

Figure 18: Price of Portfolio and Underlying Asset

The price of the portfolio x_{t-1}^P can be calculated by Equation 20 and then we get that $\bar{M}$ is 153.092. And then we can get x_α^P by using Equation 16, which is 2.7859578 when the strike price of Lamb deadweight is 426.5079. After that, we are able to find that the suitable date to buy the portfolio α is April,2017 by using the algorithm. Subsequently, basing on Equation 22, the total amount of each asset can be calculated, which are respectively 79.60784, 29.08748 and 44.39668.

Assets	Rubber	Soybean	Wheat
Price of chosen date	240.924	316.599	142.860

Due to the existence of the error, the profit we gain by using the strategy may not be equal to the payoff. The profit we gain from Wheat is 0.4089 and

the loss for another two assets are respectively 0.06241 and 0.072099, which means the profit is 0.91051 so the total profit is 42.006965. After deducting the loss from the spread option, which is 26.26786, we could get a total gain of 15.7391. Therefore it can actually successfully covers the payoff which we can see as a good performance.

Date	Rubber	Soybean	Wheat	Lamb	Portfolio	M	Xp
31-Jan-2017	253.39837	337.38102	143.77000	381.5	2.880883	132.42469	2.785958
28-Feb-2017	257.51070	344.08632	148.63000	381.575	2.831048	134.78222	
31-Mar-2017	245.67547	325.53644	142.61000	391.925	2.813929	139.28035	
30-Apr-2017	240.92404	316.59950	142.86000	413.22	2.762925	149.55889	
31-May-2017	238.28503	300.23631	149.37000	449.6	2.681799	167.64863	
30-Jun-2017	242.91610	301.80900	159.39000	501.37	2.591544	193.46385	
31-Jul-2017	244.77823	321.93715	155.71000	467.97	2.579388	181.42675	
31-Aug-2017	243.09848	321.41414	148.75000	437.6	2.670066	163.89106	
30-Sep-2017	243.80363	308.65796	150.78000	402.22	2.71746	148.01321	
31-Oct-2017	242.88554	317.26503	146.62000	386	2.727275	141.5332	
30-Nov-2017	244.65753	319.50987	143.00000	392	2.81945	139.03421	
31-Dec-2017	241.04406	316.22003	144.27000	400.24	2.740485	146.04716	
Profit	(0.12002)	(0.37947)	1.41000		$\bar{M}$	153.09202	
Profit per P	(0.06241)	-0.072099	0.40890			Strike	426.5079
Total	0.27439						
$\bar{M}*P$	42.006965					Payoff	26.26786
						Gain	15.7391

Figure 19: Calculation Results

4.5 Limitation

There are three main limitations in our trading strategy. The first one is that the strategy is based on a good result in R-square while it may hard to achieve in the multiple linear regression model, which may cause a certain error. Another one is that due to the similar asset price trend of A and P, we only consider the situation that there existed a loss in the spread option then we can gain a profit by implementing the future portfolio while if we are in the opposite situation, we would face a loss. Thirdly, the model is based on the strong correlation mathematically but not causally while the price trend could also be influenced by external factors, which may be a risky aspect of the trading strategy.

5 Conclusion

In conclusion, two methods are provided for the Stable company to hedge with the existing products to offset the risk of adverse price movements in the market. The first strategy that includes hedging with single asset is much more straightforward, which means it can be easier to understand and to apply in real life. It set a benchmark in order to help decide our trading

position. However, due to the high transaction cost, failure to fully hedge the risk and its own limitation that the derivatives of stable company's product might not be able to be traded in reality, it cannot fully satisfy our need.

We also put forward a second strategy which is more specific and mature. Basing on the general idea of trading with multiple closely related futures to hedge our risk, we first find our portfolio that are used to hedge the risk using machine learning algorithms. By considering Stable's commodities as experimental subjects and combining them with commodities from database, corresponding similar products are selected through clustering algorithms. After constructing MLR model, according to the training dataset, weights from two methods are approximately consistent. Meanwhile, R squares values as measurement of fitting are generally acceptable and are different for each possible portfolios of ten commodities respectively. Basing on this, we decided the quantity of each futures included in the portfolio, the amount of the whole portfolio as well as the timing to trade it. After a few simulations, we saw that this strategy can gave a very good performance. Therefore, it is able to hedge risks even when there is no derivative available on the market that can alone offset the risk faced by the company. However, due to the limitation of the data provided to us, we recommend financial institutions further this research with more detailed data.

References

[1] Neil D. An efficient approach for pricing spread options. *THE J. OF DERIVATIVES*, 1995.

[2] Nusa Erman, Ales Korosec, and Jana Suklan. Performance of selected agglomerative hierarchical clustering methods. *Innovative Issues and Approaches in Social Sciences*, 8:180–204, 01 2015.

[3] James K Glassman and Kevin A Hassett. Dow 36,000. *Atlantic Monthly*, 284(3):37–51, 1999.

[4] B.I. Jacobs, K.N. Levy, and M.J.P. Anson. *Market Neutral Strategies*. Frank J. Fabozzi Series. Wiley, 2005.

[5] Hong J. Kan, Hadi Kharrazi, Hsien-Yen Chang, Dave Bodycombe, Klaus Lemke, and Jonathan P. Weiner. Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. *PLOS ONE*, 14(3):1–13, 03 2019.

[6] Jacob Kogan. *Introduction to clustering large and high-dimensional data*. Cambridge University Press, 2007.

[7] Karel Kuchar, Eva Holasova, Lukas Hrboticky, Martin Rajnoha, and Radim Burget. Supervised learning in multi-agent environments using inverse point of view. In *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, pages 625–628. IEEE, 2019.

[8] Yu-Sheng Lai. Dynamic hedging with futures: a copula-based garch model with high-frequency data. *Review of Derivatives Research*, 21(3):307–329, Oct 2018.

[9] Jonathan Law. *A dictionary of finance and banking*. Oxford University Press, 2014.

[10] Rosa Lletı, M Cruz Ortiz, Luis A Sarabia, and M Sagrario Sánchez. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, 515(1):87–100, 2004.

[11] Harry Markowitz. Portfolio selection*. *The Journal of Finance*, 7(1):77–91, 1952.

[12] Atle Oglend and Frank Asche. Cyclical non-stationarity in commodity prices. *Empirical Economics*, 51(4):1465–1479, Dec 2016.

[13] John Okunev and Derek White. Do momentum-based strategies still work in foreign currency markets? *Journal of Financial and Quantitative Analysis*, 38(2):425–447, 2003.

[14] Hanna Sikacz and Przemyslaw Wolczek. Esg analysis of companies included in the respect index based on thomson reuters eikon database. *Research Papers of Wrocław University of Economics*, 2018.

[15] S. Wei, F. Wang, and D. Jiang. Sparse component analysis based on an improved ant k-means clustering algorithm for underdetermined blind source separation. In *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, pages 200–205, May 2019.

[16] Pelin Yildirim and Derya Birant. K-linkage: A new agglomerative approach for hierarchical clustering. *Advances in Electrical and Computer Engineering*, 17(4):77–89, 2017.

[17] Wei Zhang, Gongxuan Zhang, Xiaohui Chen, Yueqi Liu, Xiumin Zhou, and Junlong Zhou. Dhc: A distributed hierarchical clustering algorithm for large datasets. *Journal of Circuits, Systems and Computers*, 28(04):1950065, 2019.