Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7778521

DETECTION OF EXTREME OBSERVATIONS IN MULTIPLE MULTIVARIATE DATA USING PROJECTION TECHNIQUES: APPLICATION TO FOOD PRICE DATA

Francis Eyiah-Bediako[1], Bismark Kwao Nkansah[1*], David Kwamena Mensah[1]
1 Department of Statistics, University of Cape Coast, Cape Coast, Ghana
∗Corresponding author (bnkansah@ucc.edu.gh; 0246 723 441)

Abstract
Projection techniques such as variants of Principal Components and Outlier Displaying Components are specifically known for application in single multivariate datasets. In this paper, extensions are made of these techniques to dataset that is multiple multivariate time-dependent (MMTD) in nature. The structure of this kind of data problem is appropriately characterized to show that a single observation is a random matrix of dimensions r multiplicities by p several variables. The procedure is a two-phased approach that identifies suspect extreme observations and then examines their extent of extremeness. The application illustrates the determination of markets with extreme agricultural food commodity prices that provides useful guide for reducing levels of extreme high prices.

Key words: projection techniques, market classification, multiple multivariate data, outlier displaying component

1 Introduction

The problem of assessing extreme observations in statistical data has usually been studied in single multivariate datasets. It means that methods that are usually used in such studies have not been designed and applied in multiple multivariate (MM) datasets. For example, the Principal Components Analysis is not primarily designed as a technique for detecting extreme observations. However, since it is used as a means of constructing indices, it may be used as a preliminary measure for detecting extreme observations. The actual challenge may be highlighted if it is applied in MM datasets. Similar reservations may be expressed about the applications of techniques that are actually designed for studying extreme observations, such as the Outlier Displaying Component (ODC) (Nkansah & Gordor, 2012b; 2013). The use of such procedures in extended multivariate data would therefore require substantial modifications. These two are the projection methods that are used in this work. For the ODC, three different variants of applications will be considered. In addition to the two main techniques, we deliberately include the Generalized Principal components to buttress a point.
Since MM data has several individual multivariate datasets with varying variance-covariance structures, the overall features of such data become quite complex for determining specific extreme observations. In order to obtain a realistic effect of variations in observations on the original Outlier Displaying Component (ODC) (Gordor & Fieller, 1994), a modification of the technique has been carried out (Nkansah & Gordor, 2013). The modification enables the use of alternative measure of the mean vector in the estimation of the variance-covariance matrix in

the ODC so that results are not influenced by extreme variations in the data. The applications so far have been limited to single multivariate data.

In this study, the time-dependent principal components are used as interim techniques to obtain suspect extreme observations. The ODCs are then used to assess the extent of extremeness of the suspect observation. The resulting classifications could be a subject of further examination.

To handle MM data, an important basic problem is the appropriate characterization of individual observation in the dataset. In this study, a clear characterization is made of the general observation in such dataset at each step of the data processing and analysis. This is necessary to enable the tracking of the particular observation that is suspected to be extreme.

Examples of MM data are as follows:

(1) Prices of same set of several commodities collected from the same set of markets for several years;

(2) Input and output cost variables of the same set of companies over a number of years;

(3) Measurements on a number of economic variables for the same set of countries over a number of years;

(4) Fatality counts from a number of accident sources among various age groups for a number of locations of a region.

The number of years in Example 1 to 3, and locations in Example 4 provide the multiplicity of the multivariate data structure in the respective cases. The observation is the vector of measurements on the several variables for the market/company/country in Example 1 to 3, and age group in Example 4. In this study the term ‘‘time’’ (which may be represented by year or location) is used generically to represent the multiplicity of the data structure.

For MM data as described in the examples above, a general observation is characterized as

$$X_{s+(t-1)n,j,t}; \quad s=1,2,\ldots,n; \quad t=1,2,\ldots,r; \quad j=1,2,\ldots,p \tag{1.1}$$

obtained on p variables from n observations for each of r time points. By varying the values of j and t, it can be deduced that the single observation X_s takes the form of a random matrix given as

$$\mathbf{X}'_s = \begin{pmatrix} X_{s11} & X_{s+n,1,2} & \cdots & X_{s+(t-1)n,1,t} & \cdots & X_{s+(r-1)n,1,t} \\ X_{s21} & X_{s+n,2,2} & & X_{s+(t-1)n,2,t} & & X_{s+(r-1)n,2,t} \\ X_{s31} & X_{s+n,3,2} & & X_{s+(t-1)n,3,t} & & X_{s+(r-1)n,3,t} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ X_{sj1} & X_{s+n,j,2} & \cdots & X_{s+(t-1)n,j,t} & \cdots & X_{s+(r-1)n,j,r} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ X_{sp1} & X_{s+n,p,2} & \cdots & X_{s+(t-1)n,p,t} & \cdots & X_{s+(r-1)n,p,r} \end{pmatrix}$$

Thus, an observation in the data constitutes $r \times p$ matrix, an extension of the usual vector observation in the case of a single multivariate data. In subsequent sections, the element in Equation (1.1) and the subset data of the *k*th time point will be represented in various ways. The characterization, as well as most parts of the presentation, is made with the background of MATLAB language.

Usually, variants of GARCH modelling approach are used in studies on extremes in agricultural commodity prices, particularly to determine factors of such events (e.g., Algieri & Leccadito, 2021; van Oordt & Stork, 2021). Such approach is also used in a few studies (e.g., Abokyi & Asiedu, 2021) that have made use of rather limited portions of the same data as used in this paper.

The determination of extreme locations based on time periods (that may be non-consecutive) could be more suitably carried out using direct computational methods. The objective of this work is therefore to extend the use of the stated projection methods for single multivariate data to multiple multivariate data in order to provide a one-dimensional assessment of the extent of extremeness. The ultimate intent is to present perspectives for monitoring data with MM structure, alternative to conventional methods of spatial statistical techniques that could understandably be laborious for such data structure. It will be demonstrated that for data of this nature, multiple techniques would be required in a single study to obtain coherent results, as results based on isolated application of individual techniques could be inconsistent.

Table 1: Variations in the illustrative data

Year	Sum of Squares	Variance
1	5.2609×10^5	5845.44
2	9.4660×10^5	10517.78
3	2.0223×10^6	22470.00
4	2.9098×10^6	32331.11
5	7.5747×10^6	84163.33
Total	1.3980×10^7	30792.95

Along with the varying variance-covariance structure components of the data problem is the challenge of increasing variations over time. Table 1 shows the sample sum of squares and corresponding variance in our illustrative price data (of the type of Example 1 above) for each of five years. The variance is obtained as the trace $\text{tr}(\mathbf{S}_k)$, where $\mathbf{S}_k$ is the variance-covariance matrix of data for the *k*th year.

The table shows increasing variation in prices over the years, which may be assigned to either or both of two causes: (1) General increases in prices across almost all markets each year; (2) Increases in prices in one or a few markets each year.

The table buttresses the observation that the pattern of crop production (Ritchie & Roser, 2020), which has been on sharp increase particularly since the 2000s (except for legumes and nuts) has not resulted in a decline or stable prices over the same period, and constitutes a good motivation for studies in the area. The motivation is also driven by price projections in reports of monitoring institutions such as the FAO (2008; 2018) and OECD (2012). These reports, among

others, indicate the impact of climate change on global crop production particularly on food-insecure areas of sub-Saharan Africa.

The paper proceeds by presenting extensions of existing methodology to MM data in Section 2. Various aspects are presented in MATLAB codes. Section 3 then presents the results of application of the extended methods. Effort is made to obtain and present results that are consistent with all the various methods used. The last portion of this section gives a brief discussion of results. In Section 4, conclusion is made with relevant recommendation.

2 Description of Methods

Description of Time-dependent Displaying Techniques

The Outlier Displaying Component

For a single multivariate data $\mathbf{X}_{n\times p} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$, the Outlier Displaying Component (Gordor & Fieller, 1999) is given by

$$\boldsymbol{\beta}_{\varepsilon} = \mathbf{S}^{-1}(\mathbf{x}_{\varepsilon} - \bar{\mathbf{x}}), \tag{2.1}$$

assuming that $\mathbf{x}_{\varepsilon}$ is a known outlier, and $\mathbf{S}$ is the sum of squares and cross-product (SSCP) matrix. It is later shown (Nkansah & Gordor, 2012b) that a better display of the observation $\mathbf{x}_{\varepsilon}$ is obtained by the vector

$$\boldsymbol{\beta}_{(\varepsilon)} = \mathbf{S}_{(\varepsilon)}^{-1}(\mathbf{x}_{\varepsilon} - \bar{\mathbf{x}}_{(\varepsilon)}) \tag{2.2}$$

called the Modified One Outlier Displaying Component (M1-ODC). In the M1-ODC, $\mathbf{S}_{(\varepsilon)}$ and the vector $\bar{\mathbf{x}}_{(\varepsilon)}$ are, respectively, the corresponding matrix and vector in Equation (2.1) computed with $\mathbf{x}_{\varepsilon}$ deleted from the data.

It is necessary to point out the features of the matrix and vectors involved in the displaying component for purposes of providing extensions for this study. Originally, the SSCP matrix $\mathbf{S}$ and mean vector $\bar{\mathbf{x}}$ in Equations (2.1) and (2.2) are based on data that is single multivariate in nature. Thus, $\mathbf{S}$ and $\bar{\mathbf{x}}$ are of dimensions $p\times p$ and $p\times 1$, respectively. Now, suppose there are measurements $\mathbf{X}_{n\times p} = (\mathbf{X}_{1k}, \mathbf{X}_{2k}, \ldots, \mathbf{X}_{pk})$, $k = 1,2,\ldots,r$ for r time points (or years). This yields $(nr\times p)$ matrix of multivariate observations, and the dimension of $\bar{\mathbf{x}}$ remains unchanged. Each observation $\mathbf{x}_{\varepsilon}$ is now a matrix of dimension $(r\times p)$ given in Equation (1.1), and which is denoted by the code

$$\mathbf{X}_{p\times r}^{(\varepsilon)} = [\mathbf{X}(\varepsilon,:)'\ \mathbf{X}(\varepsilon+n,:)'\ \mathbf{X}(\varepsilon+2*n,:)'\ \ldots\ \mathbf{X}(\varepsilon+(r-1)*n,:)'] \tag{2.3}$$

The nature of the dataset presents two ways of computing the matrix, $\mathbf{S}$. It can be measured as the total SSCP based on the entire $(nr\times p)$ dataset. However, this can unduly enhance the projection of the suspect outlier in a year that has a very high variation in X_i. A fair projection could be obtained by using the pooled SSCP matrix. Thus, consider the years as constituting r groups and denote the year by a categorical variable, T. Let $T(\omega)$ be the year of observation, ω. The frequency of all year classes is n. The within-year SSCP matrix for the kth year is

$$\mathbf{W}_k = \sum_{\omega:T(\omega)=k}^{n}\left(\mathbf{X}'_{i,k}(\omega)-\bar{\mathbf{x}}_{i,k}\right)*\left(\mathbf{X}'_{i,k}(\omega)-\bar{\mathbf{x}}_{i,k}\right)'; \quad i=1, 2,\ldots, p; \quad k=1, 2, \ldots, r. \tag{2.4}$$

Dropping the index *i* for the variable, we have

$$\mathbf{W}_k = \left(\mathbf{X}'_k(\omega)-\bar{\mathbf{x}}_k(\omega)\mathbf{1}'\right)*\left(\mathbf{X}'_k(\omega)-\bar{\mathbf{x}}_k(\omega)\mathbf{1}'\right)', \tag{2.5}$$

where $\mathbf{1} = \text{ones}\ (n,\ 1)$. The pooled within-year SSCP is then given by

$$\mathbf{S}_{pooled} = \sum_{k=1}^{r}\left(\mathbf{X}'_k(\omega)-\bar{\mathbf{x}}_k(\omega)\mathbf{1}'\right)*\left(\mathbf{X}'_k(\omega)-\bar{\mathbf{x}}_k(\omega)\mathbf{1}'\right)' \tag{2.6}$$

For this study therefore, the M1-ODC will be extended to a multiple modified 1-ODC (MM1-ODC) given as

$$\boldsymbol{\beta}_{(\varepsilon)} = \mathbf{S}^{-1}_{(\varepsilon)}(\mathbf{X}^{(\varepsilon)} - \overline{\mathbf{X}}_{(\varepsilon)}) \tag{2.7}$$

Based on the pooled SSCP, Equation (2.7) may then be given as

$$\boldsymbol{\beta}_{p(\varepsilon)} = \mathbf{S}^{-1}_{pooled(\varepsilon)}(\mathbf{X}^{(\varepsilon)} - \overline{\mathbf{X}}_{(\varepsilon)}), \tag{2.8}$$

where the mean matrix in Equation (2.8) is now $(p\times r)$ matrix $\overline{\mathbf{X}}_{(\varepsilon)} = \left[\bar{\mathbf{x}}_1,\ \bar{\mathbf{x}}_2,\ \ldots,\ \bar{\mathbf{x}}_r\right]$ obtained without the suspect outlier observation, $\mathbf{X}^{(\varepsilon)}$, and $\bar{\mathbf{x}}_k$ is the mean vector for data in Year *k*. It is noted that the pooled SSCP itself is used advisedly as it may not yield results that are consistent with those based on individual group (year) data for a particular technique.

Principal Components

We shall denote the variables by $X_1, X_2, \ldots, X_p,\ (p=19)$ for convenience. The representations of X_i can be seen in Table A2. Denote the *i*th principal component (PC) of the *k*th year by y_{ik} defined as

$$y_{ik} = \sum_{j=1}^{p} a_{ikj}x_j. \tag{2.9}$$

For purposes of plausible interpretation, it is usually necessary to identify the influential indicators in the formation of y_{ik} by a large weight, a_{ij}. However, the nature of the data may not facilitate this. For example, Table A2 is the first two PCs y_{12} and y_{22}, extracted from data on Year 2 based on the variance-covariance matrix. Only x_6 is influential on y_{12}, which is as a result of high variation in that item. The second PC, y_{22}, cannot be attributed to any of the variables since $a_{2j},\ \forall j,$ are low. Since the real importance of x_j in Equation (2.9) can be influenced by its variation, we would rather obtain y_{ik} in terms of the loadings and expressed as a factor component as

$$f_{ik} = \sum_{j=1}^{p} l_{ikj} x_j ; \qquad i = 1, 2, \ldots, q ; \; k = 1, 2, \ldots, r \tag{2.10}$$

Values l_{ikj} are the loadings with $\sum_{j=1}^{p} l_{ikj}^2$ equal to the eigenvalue λ_i of f_{ik} and represents the variation in the data explained by the component. The weight in Equations (2.9) and (2.10) are connected by

$$a_{ij} = \frac{l_{ij} s_{x_j}}{\sqrt{\lambda_i}} . \tag{2.11}$$

Denote by f_{imk} the *i*th factor component score for market *m* in Year *k*. The vector of factor scores is given by

$$\mathbf{f} = \mathbf{z}_{1\times p} \mathbf{R}^{-1}_{p\times p} \mathbf{L}_{p\times q}, \tag{2.12}$$

where $\mathbf{R}^{-1}\mathbf{L}$, which we denote by $\mathbf{C}$ is the factor score coefficient matrix and **R** and **L** are the correlation and loading matrix, respectively. The vector **z** gives the standardised values of the random variables. In Equation (2.12), we may have

$$f_{imk} = \sum_{j=1}^{p} c_{ij} z_j , \tag{2.13}$$

where $z_j = \frac{x_{ij} - \bar{x}_j}{s_{x_j}}$. If $c_{ij} > 0$ and large, and $x_{ij} > \bar{x}_j$ for some $j \in \mathbf{H} = \{h_1, h_2, \ldots, h_g\}$, then f_{imk} could be high and positive. Now let the remaining items in market *m* be x_{mt}, $t \in \mathbf{X} \backslash \mathbf{H}$. Then, we can split Equation (2.13) into two components as

$$f_{imk} = \sum_{j\in\mathbf{H}} c_{ij} z_j + \sum_{t\in\mathbf{X}\backslash\mathbf{H}} c_{it} z_t \tag{2.14}$$

In the second summand in Equation (2.14), $c_{it} z_t < 0$, since $c_{it} > 0$ and $x_{it} - \bar{x}_t < 0$ for some *t*. Therefore, a high positive score reflects a market that has high prices with respect to x_j, $j \in \mathbf{H}$. Similarly, a high negative score that reflects a market with low prices may be described. We subsequently obtain the first *q* PCs for each year data and denote the resulting scores as

$$\mathbf{f} = [\mathbf{f}_{1k}, \ldots \text{sign}_{k\in I} \mathbf{f}_{ik} \ldots \mathbf{f}_{qk}] , \tag{2.15}$$

which is $(nr \times q)$ matrix. The sign of the factor score helps to determine the correct label of the suspect outlier market. For the purpose of generating composite graph on all scores simultaneously for all years, the direction of interpretation must be the same for all PCs. However, this is not the case for time-dependent component scores which is caused by inconsistency in sign. This is particularly noted for $k = 1$ on PC1. To resolve this, we define the sign of $\mathbf{f}_{ik}$ in Year *k* as follows:

$$\text{sign}_{k\in I} \mathbf{f}_{ik} = \begin{cases} -\sum_{j=1}^{p} l_{ikj} z_j , & l_{ikj} < 0 \quad \forall |l_{ikj}| > \tau \\ \sum_{j=1}^{p} l_{ikj} z_j , & l_{ikj} > 0 \quad \forall |l_{ikj}| > \tau \end{cases} \tag{2.16}$$

for the set *I* of years for which $\mathbf{f}_{ik}$ has sign inconsistency. The value of τ is a reasonably chosen loading which serves as a cut-off for determining those x_j that influence the formation of $\mathbf{f}_{ik}$. Usually $\tau = 0.5$ (Frempong et al., 2017; Nkansah, 2018). By this treatment, the resulting score signs are in line with those of other years.

However, this rule is not applicable to certain years. For those years there is the incidence of contrasting components in which some groups of indicators have high positive loadings whilst others have high negative loadings. In this case, a high positive score would mean that the market is high-priced on items that have positive loadings but low-priced on those with negative loadings. On the other hand, a high negative score could mean that the market is high-priced on items with negative loadings but low-priced on items with positive loadings.

Generalized Principal Components

In order to enhance projection of extreme observations in ordinary PCA, the generalized principal component analysis (GPCA) may be used. The GPCA is based on the eigenvectors of the product matrix $\mathbf{SS}^{*-1}$ associated with the *q* largest eigenvalues, where $\mathbf{S}$ is the usual variance-covariance matrix and $\mathbf{S}^*$ is defined as

$$\mathbf{S}^* = \frac{\sum_{j=1}^{n} K\left(\left\|\mathbf{x}_j - \mathbf{x}^*\right\|^2_{\mathbf{S}^{-1}}\right)\left(\mathbf{x}_j - \mathbf{x}^*\right)\left(\mathbf{x}_j - \mathbf{x}^*\right)'}{\sum_{j=1}^{n} K\left(\left\|\mathbf{x}_j - \mathbf{x}^*\right\|^2_{\mathbf{S}^{-1}}\right)},$$

and $\mathbf{x}^*$ is a vector of medians, $\|\mathbf{X}\|^2_{\mathbf{M}} = \mathbf{X}'\mathbf{M}\mathbf{X}$ and K is a decreasing function defined as $K(u) = \exp(-hu)$. We set $h = 0.1$ (Caussinus & Ruiz, 1990). The computation of $\mathbf{S}^*$ therefore assigns less weight to an observation that is distant from the centre of the data than those that are close. The projection pursuit (Caussinus & Ruiz-Gazen, 1993) of $\mathbf{Z}_k \times GP_i$, $i = 1, 2, \ldots, r$ of standardized data $\mathbf{Z}_k$ of the *k*th year onto the *i*th component GP_i of the product matrix $\mathbf{SS}^{*-1}$ is expected to yield clearer extreme observations. The GPCA is included in this study particularly to determine how it yields other extreme observations than those obtained from PCA. For both the PCA and GPCA, we label an observation as having a level of extremeness based on the following guide on the index value *C* of its projection.

Table 2: Interpretation of component values

Limits of Index	Representation
$C \geq 3$	Extremely High
$2 \leq C < 3$	High
$-2 \leq C < 2$	Moderate
$-3 < C < -2$	Low
$C \leq 3$	Extremely Low

The cut-off values for classification based on principal components is ultimately informed by what could also appropriately constitute an extreme value on the basis of other projection methods.

Table 3 presents the summary and interpretation of basic notations that have been used frequently in the paper.

Table 3: Summary of basic notations and interpretations

Notation	Interpretation
$\mathbf{X}$	Initial $n \times p$ data matrix
$\mathbf{X}_k$	$n \times p$ data matrix for the *k*th year
$\mathbf{x}_\varepsilon$	$p \times 1$ data vector for observation ε in a single multivariate data
$\bar{\mathbf{x}}_{(\varepsilon)}$	$p \times 1$ mean vector with observation ε deleted from data
$\mathbf{X}^{(\varepsilon)}$	$r \times p$ data matrix for observation ε in MM data
$\mathbf{S}$	Sum of squares cross-product (SSCP) matrix in ODC, but may also represent variance-covariance matrix in principal components and measures of statistical distance
$\mathbf{S}^*$	Weighted variance-covariance matrix used for computing GPCA
$\mathbf{S}_{(\varepsilon)}$	SSCP matrix with observation ε deleted from data
$\boldsymbol{\beta}_\varepsilon$	Original ODC based on observation ε as suspect outlier
$\boldsymbol{\beta}_{(\varepsilon)}$	Modified ODC (M1-ODC) based on observation ε as suspect outlier deleted from data
$\boldsymbol{\beta}_{p(\varepsilon)}$	M1–ODC based on pooled SSCP matrix with observation ε deleted
$\mathbf{S}_{pooled(\varepsilon)}$	Pooled SSCP matrix with observation ε deleted
f_{imk}	The *i*th component score for market *m* in Year *k*.
$U(\mathbf{x}_\varepsilon; \bar{\mathbf{x}}, \mathbf{S})$	Statistical distance of observation ε from the mean $\bar{\mathbf{x}}$ based on variance matrix, $\mathbf{S}$
$\mathbf{X}(\varepsilon + (t-1) * n, :)$	Data vector in (or row of) $\mathbf{X}$ for observation ε in the *t*th year

Assessment of Significance of Extremeness and Implementation Procedure

It has been shown (Nkansah & Gordor, 2013) that based on $\mathbf{S}_{(\varepsilon)}^{-1}$ and $\bar{\mathbf{x}}_{(\varepsilon)}$, with $\mathbf{x}_\varepsilon$ deleted, the distance $U(\mathbf{x}_\varepsilon; \bar{\mathbf{x}}, \mathbf{S}) = (\mathbf{x}_\varepsilon - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_\varepsilon - \bar{\mathbf{x}})$ is related to $U(\mathbf{x}_{(\varepsilon)}; \bar{\mathbf{x}}_{(\varepsilon)}, \mathbf{S}_{(\varepsilon)})$ by the relation

$$\frac{U(\mathbf{x}_\varepsilon; \bar{\mathbf{x}}, \mathbf{S})}{U(\mathbf{x}_{(\varepsilon)}; \bar{\mathbf{x}}_{(\varepsilon)}, \mathbf{S}_{(\varepsilon)})} = \left(\frac{n-1}{n}\right)^2 \left[1 - \lambda U(\mathbf{x}_{(\varepsilon)}; \bar{\mathbf{x}}_{(\varepsilon)}, \mathbf{S}_{(\varepsilon)})\right],$$

where $\lambda = \frac{n-1}{n} \cdot \frac{1}{1+tr\left(\mathbf{A}_I \mathbf{S}_{(\varepsilon)}^{-1}\right)}$, and $\mathbf{A}_I = \frac{n}{n-1}(\mathbf{x}_\varepsilon - \bar{\mathbf{x}})(\mathbf{x}_\varepsilon - \bar{\mathbf{x}})'$ (Nkansah & Gordor, 2013; Barnett & Lewis, 1994). Since $U(\mathbf{x}_{(\varepsilon)}; \bar{\mathbf{x}}_{(\varepsilon)}, \mathbf{S}_{(\varepsilon)}) > U(\mathbf{x}_\varepsilon; \bar{\mathbf{x}}, \mathbf{S})$, it means that if an observation is found to be significantly extreme on the pooled reduced SSCP, $\mathbf{S}_{(\varepsilon)}$, then it is necessarily significantly extreme on the pooled $\mathbf{S}$. It is therefore reliable to determine the significance of extreme observation by the distinctness of its projection based on $\mathbf{S}_{(\varepsilon)}$.

The main steps for executing the procedures in the study are summarized in Figure 1.

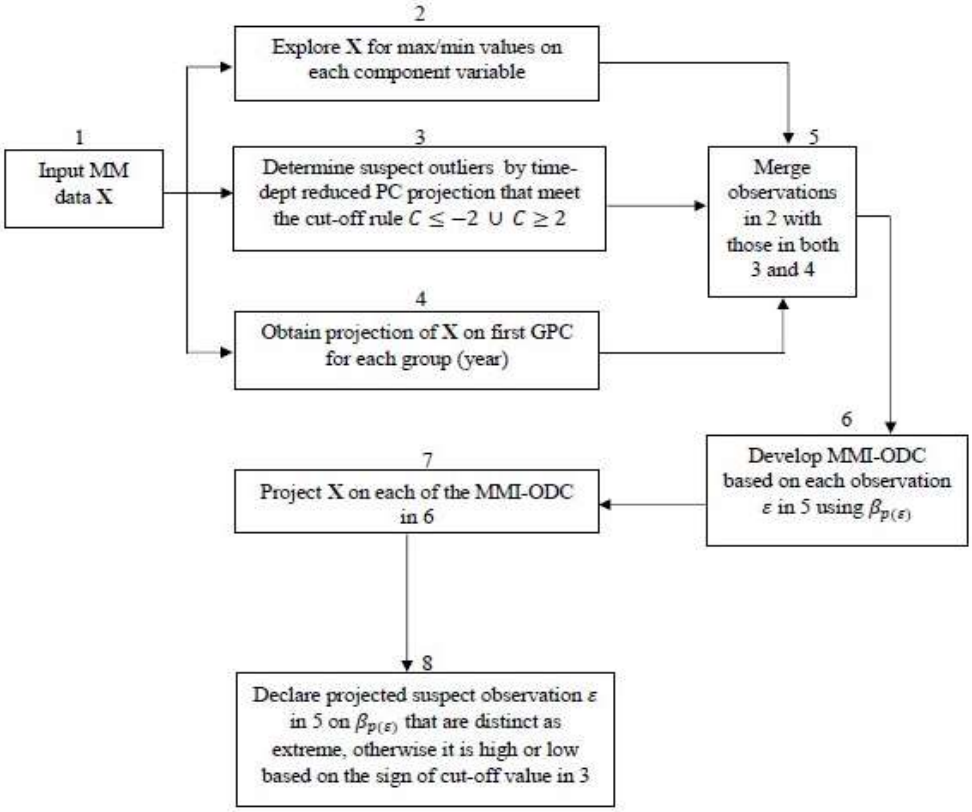

Figure 1: Flowchart for execution of the methods

3 Application to Price of Local Food Price Data

The data used is price data obtained from 91 markets on 19 local food items that covers eight main categorizations of local foodstuff, namely: Cereals, Roots and Tubers, Vegetables, Pulses, Fish, Spices, Oil and Fruits. The data covers five non-consecutive years from 2008 to 2015, and are obtained from the Statistical, Research and Information Directorate (SRID) of the Ministry of Food and Agriculture (MoFA), Accra, Ghana. (See Appendix for distribution of markets.)

Preliminary Detection of Extreme-Priced Markets

The multiple nature of the multivariate data requires consistency checks. One way is to present useful background information of the data so that the eventual results would be one that agree with the background. Relevant aspects of descriptive statistics of the prices of food items are summarized in Table 4. In the table, some important observations could be made regarding

markets (e.g., 17, 68, 65) that are frequently associated with the minimum and/or maximum prices of respective food items.

However, a market cannot be labeled yet as extreme priced merely on the basis of descriptive statistics on a few individual commodity variables. Similarly, there could be markets that are not identified by the basic statistics that could emerge as important extreme priced markets.

In Table 4, it may be easy to identify market 65 as extreme since it has the highest number of maximum prices of items (five of them) but has a minimum price in one. The performance of this market may be examined from another perspective. Figure 2 is a display of scores $\mathbf{f}_{ik}, (i = 1, 2, 3;\ k = 1, 2, \ldots, 5)$ for the first three principal component factors for each of the five years. Displays of scores for components beyond three do not show clear extremes. In Figure 2, we can identify observation 65 as suspect extreme (high) in Year 1 and extreme (low) in Year 4 on PC1. On PC2, it is quite high in Year 4, and does not however feature on PC 3.

Table 4: Extreme Priced Markets Based on Descriptive Statistics

Commodities	Minimum Priced Market	Maximum Priced Market
Root and Tubers		
Yam white	29, 44, 65, 63, 60	1, 3, 52, 14
Cassava	8, 17, 19, 16	49, 65, 51, 89
Plantain (Apentu)	17, 16, 34, 14	59, 71, 46, 65
Gari	16, 44, 19, 77, 4	9, 21, 86, 10
Vegetables		
Tomato	12, 17, 83, 11	48, 60, 70
Garden egg	87, 58, 13, 63	65, 55, 51, 48
Cereal		
Local Rice	6, 17, 19	87, 53, 1, 7
Imported Rice	77, 34, 57	41, 21
Maize	68	77, 10, 35, 53
Oil		
Palm oil	87, 29, 21, 26	63, 79, 40
Fruit		
Orange	39, 3, 63, 6	65, 69
Banana	39, 17, 10, 20	70, 69, 56
Fishes		
Smoked herring	29, 50, 76	69, 55, 67, 71
Koobi	82, 60, 12	23, 22, 78, 7
Egg	19, 40, 1, 6	**65**, 61, 55
Spices		
Dried pepper	12, 19, 59, 60	43, 69, 89
Onion	68, 86, 63, 21	46, 41, 11, 56
Pulses		

See Figure A1 for geographical distribution of market numbers

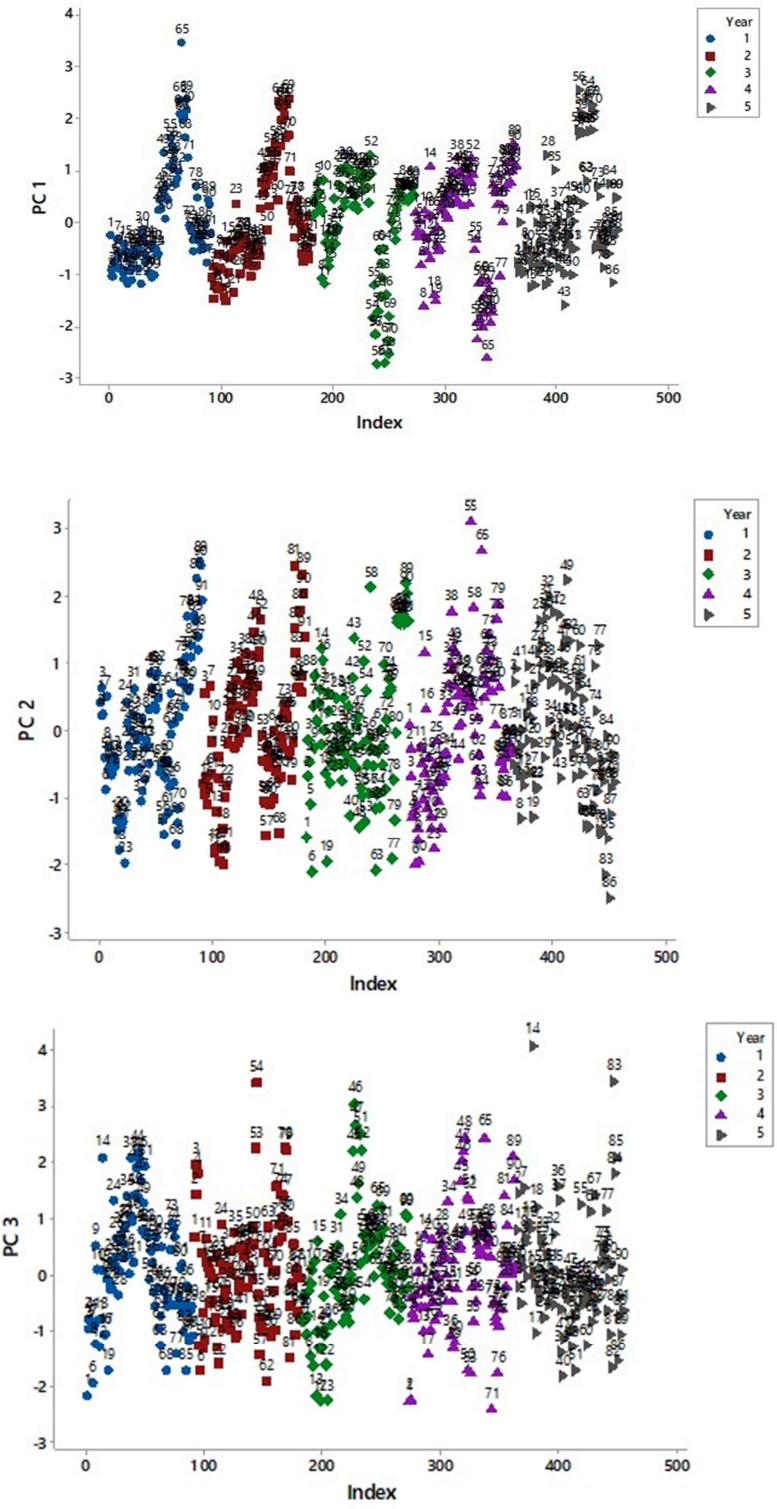

Figure 2: Scatterplots of scores of Components 1, 2 and 3 for all years

What constitutes the data problem is how to eventually categorize such an observation. Examination of all projections on the first five $(q=5)$ factors that meet the cut-off values $(C \le -2 \cup C \ge 2))$ in addition to most frequent markets in Table 4 gives the set of typical suspect outliers as $L=\{65, 69, 13, 68, 54, 89, 90, 17, 56, 46, 14, 83, 55, 47\}$. Denote this set simply by $L=\{l_1, l_2, \ldots, l_v\}$. It is worth noting that 17, which is most consistently low-priced, does not feature as extreme in any of the plots. It is observed that the eigenvalues of the Generalized PCs are not too distinctly different, and the resulting suspect extreme observations are almost the same as those based on PCA.

Detection of Extreme-Priced Markets Using Displaying Components

Denote the extracted data for the *k*th year with code $\mathbf{X}_k = \mathbf{X}\left((k-1)*n+1:k*n,\ 1:p\right)$ from the entire dataset **X**. The projection of an observation is obtained by $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}_{(\varepsilon)}$. In particular, the projection of $\mathbf{X}^{(\varepsilon)}$ from $\bar{\mathbf{x}}$ is the element of $\mathbf{Y}$ given by

$$\mathbf{y}_{p\times r}^{(\varepsilon)} = \left[\mathbf{Y}(\varepsilon, 1)\ \ \mathbf{Y}(\varepsilon+91, 2)\ \ \mathbf{Y}(\varepsilon+2*91, 3)\ \ \ldots\ \ \mathbf{Y}(\varepsilon+(r-1)*91, r)\right]$$

The projection of all observations $\varepsilon \in L$ based on the total sum of squares and cross-product (SSCP) and the pooled SSCP matrix are given in Tables 5 and 6, respectively. Markets 65 in Year 4, and 14 and 83 in Year 5 show extremely large distances indicating high price levels in those years. We also note that market 17 is very consistent among the lowest priced each year. As noted in Section 1, the projected values in Table 5 are highly affected in those years with wide variations.

Table 5: Projected values from general mean of suspect outlying markets based on total SSCP matrix

		Year				
No.	ε	1	2	3	4	5
1	65	13.9311	3.0370	13.8802	**263.9890**	47.7801
2	69	6.6609	23.7912	11.0015	13.9365	58.8302
3	13	1.9278	1.3535	19.6604	16.1751	27.4488
4	68	2.7572	21.6740	7.7632	13.0138	58.9821
5	54	3.9404	7.6739	16.7003	11.0493	37.3578
6	89	4.2535	3.7267	15.0472	25.1642	23.7623
7	90	4.7164	3.0395	14.4356	20.6649	49.0961
8	**17**	**1.9614**	**0.6981**	**8.4387**	**14.5450**	**55.0455**
9	56	5.2528	3.0540	24.3344	30.9917	64.3164
10	46	1.8222	6.4996	17.8794	23.3667	20.0537
11	14	4.9814	2.6910	16.4266	52.9580	**138.6904**
12	83	2.1387	1.8405	9.2228	3.4933	**114.7280**

Table 6: Projected values from general mean of suspect outlying markets based on pooled SSCP matrix

		Year				
No.	ε	1	2	3	4	5
1	65	0.0889	0.0147	0.0488	**0.5605**	0.1112
2	69	0.0320	0.0538	0.0076	-0.0096	0.1256
3	13	0.0011	-0.0058	0.0252	0.0040	-0.0263
4	68	0.0047	0.0402	**-0.0093**	**-0.0251**	0.0818
5	54	0.0085	0.0299	0.0276	0.0120	0.0411
6	89	0.0282	0.0273	0.0542	0.0829	-0.0465
7	90	0.0224	0.0222	0.0566	0.0743	0.1814
8	**17**	**-0.0003**	**-0.0035**	**-0.0205**	**-0.0227**	**0.0642**
9	56	0.0309	0.0221	0.0508	0.0721	0.2349
10	46	0.0132	0.0191	0.0784	0.0719	0.0201
11	14	0.0174	0.0054	0.0344	0.1245	0.3768
12	83	0.0109	0.0093	0.0155	-0.0005	0.3061
13	55	0.0151	0.0219	0.0624	0.1424	0.2570
14	47	0.0142	0.0175	0.0699	0.0818	0.1005

The large distances are reduced by the pooling process in Table 6 which introduces negative projections as indications of much lower prices in those markets than the mean prices. It should be noted that the projections are measures of distances of the observation from the indicated centre of the data.

The observation matrix $\mathbf{X}^{(\varepsilon)}$ may not be extreme in all the r values in the vector $\mathbf{Y}^{(\varepsilon)}$. Suppose it is extreme particularly in the value $\mathbf{Y}(\varepsilon+(k-1))*n,\ k)$ in the kth year. This value is then projected among all values in that year using the MM1-ODC given in Equation (2.7).

Now, we exclude the outlying MMTD observation from the data over all years by excluding every nth observation beginning from $\mathbf{X}_{\varepsilon}$ by the following set of r codes:

$$\mathbf{T}=\mathbf{X};\quad \mathbf{T}(\varepsilon+j*n-j\ ,\ :)=[\ ];\qquad j=0,1,..,(r-1)$$

The data is now reduced to $(n-r)$ in size with each year containing $(n-1)$ data points. The reduced data for the kth year after deleting $\mathbf{X}^{(\varepsilon)}$ is

$$\mathbf{X}_{k\backslash(\varepsilon)}=\mathbf{T}\left((k-1)*(n-1)+1\ :\ k*(n-1),\ \ 1:p\right) \tag{3.1}$$

The mean vector of this reduced dataset is $\overline{\mathbf{X}}_{k\backslash(\varepsilon)}=\left[\overline{\mathbf{X}}_{1\backslash(\varepsilon)},\ \overline{\mathbf{X}}_{2\backslash(\varepsilon)},\ ...,\ \overline{\mathbf{X}}_{r\backslash(\varepsilon)}\right]$, which is a $(p\times 1)$ mean vector for the reduced data of size $(n-1)$ in Year k and (\) means set-minus. The SSCP matrix of data in Equation (3.1) is

$$\mathbf{W}_{k\backslash(\varepsilon)}=\left(\mathbf{X}'_{k\backslash(\varepsilon)}-\overline{\mathbf{x}}_{k\backslash(\varepsilon)}\mathbf{1}'\right)*\left(\mathbf{X}'_{k\backslash(\varepsilon)}-\overline{\mathbf{x}}_{k\backslash(\varepsilon)}\mathbf{1}'\right)' \tag{3.2}$$

The corresponding pooled SSCP matrix similar to that in Equation (2.6) is constructed as

$$\mathbf{S}_{pooled\,(\varepsilon)} = \sum_{j=1}^{r} \left(\mathbf{X}'_{j\setminus(\varepsilon)} - \bar{\mathbf{x}}_{j\setminus(\varepsilon)}\mathbf{1}'\right) * \left(\mathbf{X}'_{j\setminus(\varepsilon)} - \bar{\mathbf{x}}_{j\setminus(\varepsilon)}\mathbf{1}'\right)' \tag{3.3}$$

The projected values of suspect observations in set L is obtained by $\mathbf{Q} = \mathbf{X}\boldsymbol{\beta}_{p(\varepsilon)}$. The projection of $\mathbf{X}^{(\varepsilon)}$ from $\bar{\mathbf{x}}_{(\varepsilon)}$ are the elements of $\mathbf{Q}$ given by

$$\mathbf{Q}^{(\varepsilon)}_{p\times r} = \left[\mathbf{Q}(\varepsilon,\ 1)\ \mathbf{Q}(\varepsilon + n,\ 2)\ \mathbf{Q}(\varepsilon + 2*n,\ 3)\ \ldots\ \mathbf{Q}(\varepsilon + (r-1)*n,\ r)\right]$$

Now, for each suspect outlier, we examine the corresponding elements $\mathbf{Q}^{(L)}$ which are given in Table 7. From the table, market 17 is low-priced in general. Another low-priced market is 68. However, the significance of extremeness of a suspect market is better appreciated in a plot along with the projection of all other observations. The multiple one-dimensional plots of projected values using $\varepsilon = 65, 17$ as suspect extremes are given in Figures 3 and 4, respectively. Figure 3 shows that the spread of prices in Year 4 is wider than any of the other years compared on the same scale. It identifies market 65 as the most outlying in that year but not extreme in the other years. Year 2 produced the least varied prices. The actual spread in each year may be ascertained by separate plots.

Table 7: Projected values from $\bar{\mathbf{x}}_{(\varepsilon)}$ of suspect outlying markets based on pooled reduced SSCP matrix

		Year				
No.	ε	1	2	3	4	5
1	65	0.0967	0.0161	0.0540	**1.4144**	0.1443
2	69	0.0341	0.0573	0.0090	-0.0085	0.1490
3	13	0.0010	-0.0061	0.0261	0.0039	-0.0293
4	68	0.0044	0.0422	-0.0107	-0.0267	0.0928
5	54	0.0089	0.0315	0.0287	0.0124	0.0460
6	89	0.0291	0.0281	0.0579	0.0900	-0.0471
7	90	0.0232	0.0232	0.0624	0.0837	0.2093
8	**17**	**-0.0007**	**-0.0044**	**-0.0225**	**-0.0257**	**0.0704**
9	56	0.0329	0.0234	0.0609	0.0789	0.2821
10	46	0.0136	0.0198	0.0831	0.0773	0.0224
11	14	0.0188	0.0056	0.0385	**0.1574**	**0.5466**
12	83	0.0115	0.0096	0.0155	-0.0010	0.3990
13	55	0.0156	0.0225	0.0799	0.1586	0.3415
14	47	0.0145	0.0181	0.0750	0.0891	0.1123

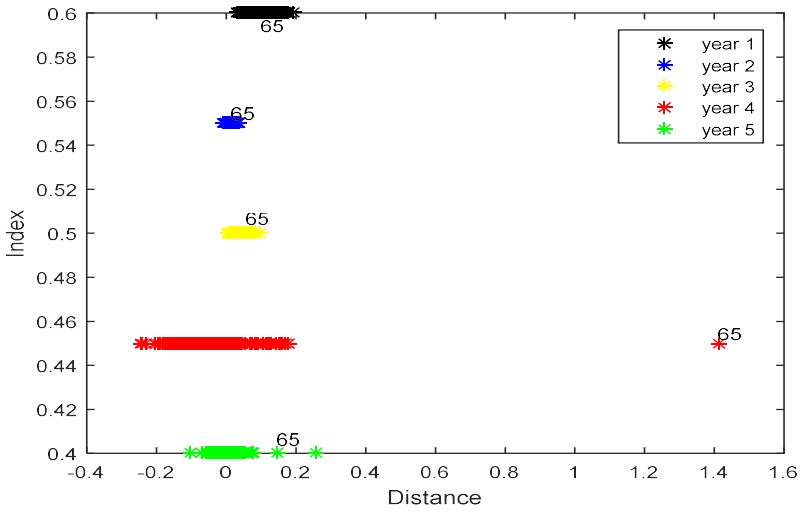

Figure 3: Multivariate projection of prices for all years using 65 as suspect outlying market

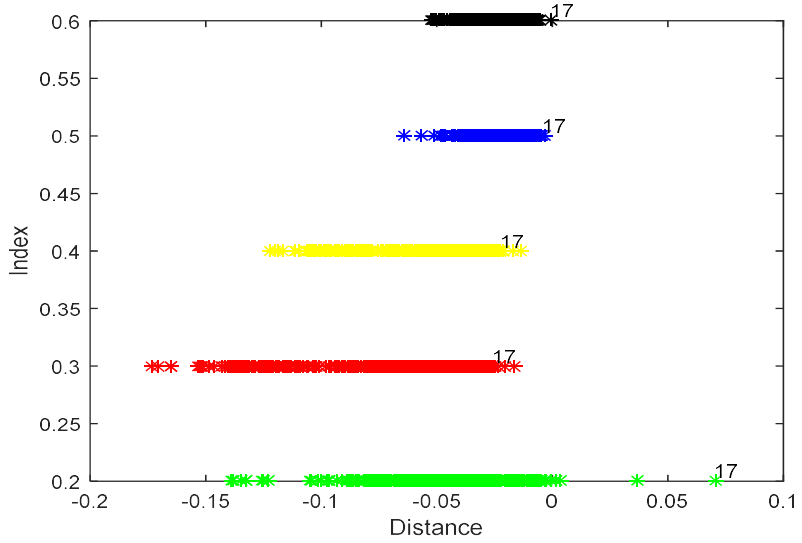

Figure 4: Multivariate projection of prices for all years using 17 as suspect outlying market

Figure 4 shows that price levels are more spread out for all years based on market 17 as suspect outlier. This means that prices in 17 are substantially lower than those of all other markets but does not constitute an extreme low priced market as it does not stick out clearly from the rest (even though it has minimum prices for several items). Graphical illustrations are given for only these two observations as they constitute the only identified typically extreme observations in the data that are worth examining for significance.

Discussion

A few of the findings are in line with results in the literature. In particular, the study has revealed that one of the most low-priced markets is market 68. This market is found to be consistently the lowest priced in Maize over the entire study period. Gage et al. (2012) rather report that most of the Maize reported in the study area is produced in the northern part of the country. This study has identified the specific location where the commodity may be obtained. It is also

found that Maize is a major influential indicator of the dimensions that have been determined in the form of time-dependent principal components (see Table A1). This means that a major component of the expenditure of consumers of local food items is taken by Maize. The relative importance of Maize, which is buttressed by similar findings in the literature, is not evident in Year 5 (2015). This observation is interestingly consistent with the break in the movement in prices of all staple products reported by Cudjoe et al. (2008). It is observed that dominant dimensions that influence the formation of components for preliminary detection of suspect extreme observations are Roots and Tubers, and Fishes and Fruits. This could be useful information for reducing levels of extreme high prices.

It is worth noting that an observation is not declared extreme simply on the basis of recording the lowest or highest values over a number of time periods. One of the typical examples of such observations is market 21. A major step in the procedure ensures that the direction of interpretation is the same for all time-dependent PCs. This requires some amount of interaction time with the data.

4 Conclusion

Various projection techniques have been explained and used to illustrate the extent of extremeness of MM observation. A key step is the characterization of MM observation as a matrix, rather than a vector for which notable projection methods are usually noted for. The structure of these techniques has therefore been appropriately extended and are applied to identify an observation that is extreme over specific time periods within the range of the data coverage. The two main approaches adopted in this work could be fused into one automated procedure to considerably reduce execution time for the proposed method.

References

1. Abokyi, E., & Asiedu, K. F. (2021). Agricultural policy and commodity price statbilization in Ghana: The role of buffer stockholding operations. *African Journal of Agricultural and Resource Economics*, **16**(4), 370 – 387.

2. Algieri, B., & Leccadito, A. (2021). Extreme price moves: an INGARCH approach to model coexceedances in commodity markets. *European Review of Agricultural Economics*, **48**(4), 878 – 914.

3. Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data*. (3rd Ed.). New York: John Wiley and Sons.

4. Caussinus, H., & Ruiz, A. (1990). Interesting projections of multidimensional data by means of generalized principal component analyses. In Compstat (pp. 121-126). Springer Int.

5. Caussinus, H., & Ruiz-Gazen, A. (1993). Projection Pursuit and Generalized Principal Component Analyses, In *New Directions in Statistical Data Analysis and Robustness*, Eds. Morgenthaler S. et al., Birkhäuse Verlag, Basel Boston Berlin, 35-46.

6. Cudjoe G., Breisinger, C., & Diao, X. (2008). Local Impacts of a Global Crisis: Food Price Transmission and Poverty Impacts in Ghana. Local Impacts of a Global Crisis, **3**, 249-269.

7. FAO (2008). Soaring Food Prices: Facts, Perspectives, Impacts and Actions required. Paper presented at High level conference on World Food Security: The Challenges of Climatic and Bioenergy, Food and Agriculture Organization, Rome, 3-5 June.

8. FAO (2018). The future of food and agriculture —Alternative pathways to 2050. Rome. 224 pp. Licence: CC BY-NC-SA 3.0 IGO
9. Gage, D., Bangnikon, J., Abeka-Afari, H., Hanif, C., Addaquay, J., Antwi, V., & Hale, A. (2012). The market for Maize, Rice, Soy and warehousing in Northern Ghana. United States Agency for International Development. Enabling Agricultural Trade (EAT).
10. Gordor, B. K., & Fieller, N. R. J. (1999). How to display an outlier in multivariate datasets. *Journal of Applied Sciences & Technology*, **4**(2) 22
11. Nkansah, B. K., & Gordor, B. K. (2013). Discordancy in Reduced Dimensions of Outliers in High-Dimensional Datasets: Application of Updating Formula. *American Journal of Theoretical and Applied Statistics*, **2**(2), 29 - 37.
12. Nkansah, B. K., & Gordor, B. K. (2012b). On the One-Outlier Displaying Component. *Journal of Informatics and Mathematical Sciences*, **4**(2), 229 - 239.
13. OECD (2012). Rising Food Prices: Causes and Consequences, Working paper, Organization for Economic Co-operation and Development.
14. Ritchie, H., & Roser, M. (2020). Agricultural Production. *Our World in Data*
15. van Oordt, M. R. C., Stork, P. A., & de Vries, C. G. (2021). On agricultural commodities' extreme price risk. *Extremes*, **24**(3), 531 – 563.

Appendix

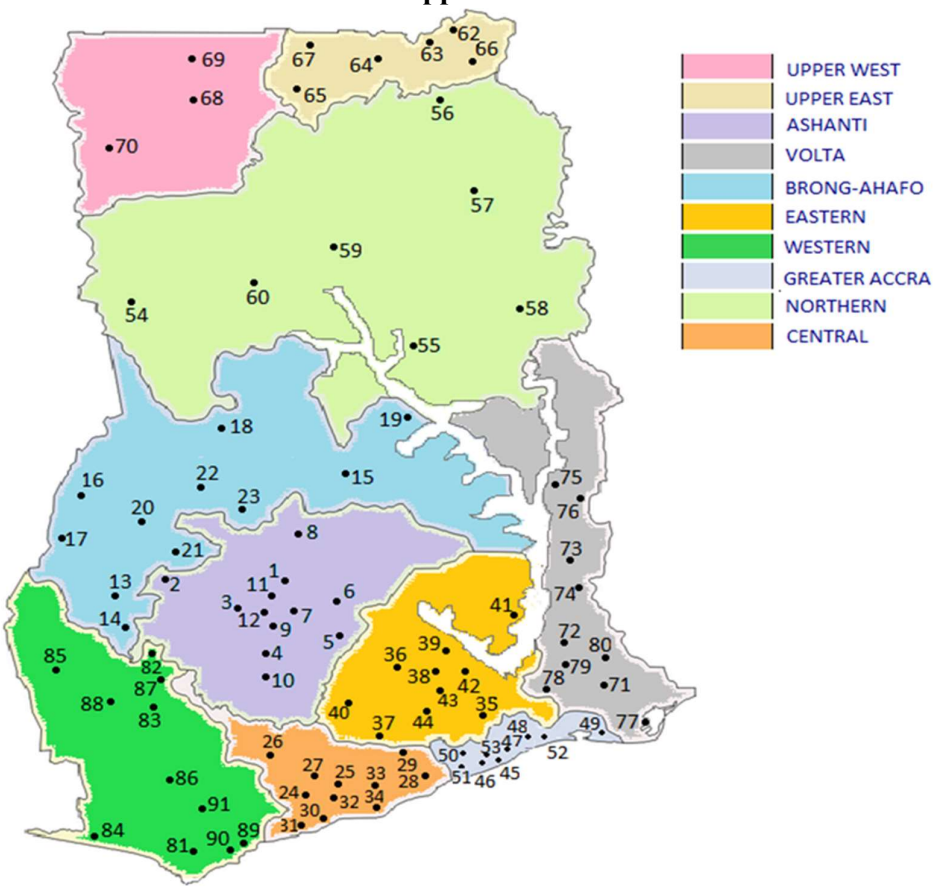

Figure A1: Map of Ghana showing distribution of study markets in dots

Table A1: Labels of Extracted Principal Components

Years	PC1	PC2	PC3	PC4	PC5
Year 1	Roots and Tubers, Garden Eggs, Fruits and Fishes,	Cereals, Yam, Tomato, Pulses	Spices	Gari, Oil	Imp Rice
Year 2	Roots and Tubers, Fruits, Fishes, Oil	Maize, Tomato Yam, Gari, Spices, Pulses	Pulse, Local Rice	Dry Pepper, Imp Rice, Oil	Garden Eggs
Year 3	Maize, Yam, Vegetable, /Fruits, Fishes	Spices, Pulses	Roots and Tubers, Vegetables, Egg	Pulse, Oil, Local Rice, Koobi	Cereal, Gari
Year 4	Maize, Yam, Spices, Vegetables, Pulses	Oil, Egg	Root and tubers, Garden Eggs	Onion, Gari, Local Rice	Fishes

Year 5	Roots and Tubers, Fruits, Fishes	Tomato, Onion	White Cowpea, Fishes, Cereals	Dry Pepper, Gari, Pulses	Oil, Egg