Cluster based area yield scheme for crop insurance policy in Java

R Haryastuti¹, M N Aidi¹, S M Pasaribu², I M Sumertajaya¹, V A Sutomo³, D Kusumaningrum³ and R Anisa¹

¹Department of Statistics, IPB University, Bogor West Java Indonesia
²Indonesian Center for Agriculture SocioEconomic and Policy Studies, Ministry of Agriculture Indonesia, Bogor West Java Indonesia
³School of Applied Science Technology Engineering & Mathematics, Prasetiya Mulya University, BSD Tangerang Banten Indonesia

E-mail: rharyastuti197@gmail.com

Abstract. Zone based area yield scheme becomes a potential in improving the development of area yield index for crop insurance policy in Indonesia. The characteristics of productivity varies among municipalities in Indonesia. Therefore, it will be challenging to apply area yield index policy without developing clusters, especially in Java which is considered as one of the central productivities of paddy in Indonesia. Clustering the productivity is one of the solutions to reduce the heterogeneous basis risk within municipalities of a province. Clustering areas is recommended in order to determine accurate critical yield index leading to a more precise ratemaking in the process of developing area yield index policy. In order to define the clusters, we have not only considered average and volatility of productivity over the past ten years, but also integrated the spatial effect of municipalities. Ward-like hierarchical method shows that clustering of municipalities within each province provides better homogeneity. It has been found that there are 12 clusters in Java tend to have low (8.33%), middle (50%), and high (41.67%) productivity. Therefore, we suggest that the Ministry of Agriculture design the area yield index policy based on these clusters and each cluster will have different critical productivity index and adjusted premium.

1. Introduction
Agriculture insurance has been introduced in Indonesia by Ministry of Agriculture (MoA) since the year 2012. MoA appointed Jasindo as a state insurance company to conduct indemnity subsidized crop insurance policy, that was indemnity-based crop policy or also known as Multi-Peril Crop Insurance (MPCI). Indemnity payments, which covers crop failures caused by floods, droughts, pests, and disease [1], are triggered by losses measured at the farm level. However, this policy carried out some disadvantages, e.g. high risk of moral hazard and adverse selection, high administrative cost, and low quality of human resources.

Sutomo et al. [2] recommended Group Risk Plan (GRP), the U.S. area yield insurance [3], could be considered as an alternative crop insurance policy in Indonesia. Nevertheless, Sutomo et al. could not obtain the precise critical yield index per group because the land area is heterogeneous. Kusumaningrum et al. [4] also conducted analytical and simulaton approach to develop alternative area yield policy in regency/ municipality level. Furthermore, it is suggested
to take consideration other area levels that could reduce the basis risks and apply a more effective policy.

Although it has been stated that a reduction in area size will improve the risk-reducing effectiveness in Indonesia, no alternative has been found to the area or zone boundaries for the area yield scheme. The characteristics of paddy productivity varies among municipalities in Indonesia, but generally it still shows similar pattern between neighbors. Therefore, identifying the different zone and classifying/assigning each municipality into appropriate zone [5] is the thrust of this research and will be challenging.

In this study, the zone-based area yield (Zone-based GRP in term of [5]) will be applied to mitigate previously mentioned risks in Java which is considered as one of the central productivities of paddy in Indonesia. The zone is referred to as a geographical region covering regencies/municipalities, each of which consist of all farms with similar yields [5]. The zones to be compared are province level and new cluster based or sub province level that is lower disaggregate of province. The province based only utilizes the administrative boundaries of each provinces. While the new cluster based will be adjusted by Ward-like hierarchical method which also considers the spatial dependencies. Historical data over past ten years is used in zone identification, i.e. average and volatility of paddy productivity.

2. Methodology

2.1. Data

The data used to identify the zone is the average of paddy productivity (yield per hectare) by regency/municipality which becomes a new potential used to obtain the critical yield index of the zone-based area yield index. It is taken from Agricultural Statistics Database of MoA since the year 2007 to 2018. In this study we focus to analyze provinces in Java Island (Table 1) because it is considered as one of the central productivities of paddy in Indonesia.

| Province     | Number of regency/ municipality |
|--------------|---------------------------------|
| West Java    | 27                              |
| Central Java | 35                              |
| DI Yogyakarta| 5                               |
| East Java    | 38                              |
| Banten       | 8                               |
| **Total**    | **113**                         |

Moreover, we also investigate the return volatility of yield rate over past ten years in each regency/municipality.

2.2. Ward-like hierarchical clustering

Chavent et al. [6] proposed a Ward-like hierarchical clustering including spatial/geographical constraints. This algorithm uses two dissimilarity matrices $D_0$ and $D_1$ and a mixing parameter $\alpha \in [0, 1]$. The first matrix $D_0$ gives the dissimilarities in the feature space which uses Euclidean distance. The second matrix $D_1$ gives the dissimilarities in the constraint space containing the geographical distances between $n$ observations. A mixing parameter $\alpha \in [0, 1]$ controls the weight of the constraint in the quality solutions. It allows the users to set the importance of each
dissimilarity matrix in the clustering procedure. In this study, we assign the \( \alpha \) of 0.5 assuming that both \( D_0 \) and \( D_1 \) are equally important.

The spirit of building a hierarchy of clusters is coming since we are officially in the land of unsupervised learning where we need to figure out patterns without a set of outcome. To obtain a new partition \( \mathcal{P}_K = (C_1, C_2, \ldots, C_K) \) in \( K \) clusters from a given partition \( \mathcal{P}_{(K+1)} \) in \( K+1 \) clusters, the idea is to aggregate the two clusters \( \mathcal{A} \) and \( \mathcal{B} \) of \( \mathcal{P}_{(K+1)} \) such that the new partition has minimum within-cluster inertia (heterogeneity, variance), that is:

\[
\arg \min_{\mathcal{A}, \mathcal{B} \in \mathcal{P}_{(K+1)}} W(\mathcal{P}_K) = \arg \min_{\mathcal{A}, \mathcal{B} \in \mathcal{P}_{(K+1)}} \sum_{k=1}^{K} I(C_k)
\]

\[
= \arg \min_{\mathcal{A}, \mathcal{B} \in \mathcal{P}_{(K+1)}} \sum_{k=1}^{K} \left( \alpha \sum_{i \in C_k} \sum_{j \in C_k} w_i w_j d_{0,ij}^2 + (1 - \alpha) \sum_{i \in C_k} \sum_{j \in C_k} w_i w_j d_{1,ij}^2 \right)
\]

where \( \mu_k = \sum_{i \in C_k} w_i \) is the weight of \( C_k \). The smaller the pseudo-inertia \( I(C_k) \) is, the more homogeneous are the observations belonging to the cluster \( C_k \). This type of clustering is also known as additive or agglomerative hierarchical clustering.

To get the number of clusters for Ward hierarchical, the result can be showed as a dendrogram. A dendrogram is a tree-like diagram that records the sequences of merges or splits. The height, or vertical line, gives the values of pseudo-inertia at which mergers occur [7]. Accordingly, we set the threshold that cuts the tallest height, which means the fathest distance between clusters.

### 2.3. Analysis procedure

A more detailed explanation of algorithm used to select zone which includes clusters of regencies/ municipalities are:

(i) Calculate the average and volatility of yield rate in 2007-2018 from each of 113 regencies/ municipalities. The functions we used can be written as:

\[
\bar{X} = \frac{\sum_{t} X_t}{12}
\]

\[
r_t = \frac{X_t - X_{t-1}}{X_{t-1}}
\]

\[
v = \sqrt{p} \sqrt{\frac{\sum_{t} (r_t - \bar{r})^2}{p}}
\]

where \( t = 2007, 2008, \ldots, 2018 \); \( X_t \) denotes the yield rate at t-time; \( \bar{X} \) and \( v \) is the average and volatility of yield rate in each regency/ municipality respectively; \( r_t \) is the return and \( \bar{r} \) is average of yield rate; \( p \) is period of returns. In this case the data is annually provided for each region, then \( p \) is equal to 1.

(ii) Check the spatial weights that will be used in Ward-like hierarchical clustering analysis. This step compares dissimilarities in the constraint space \( D_1 \) whether it is better to use Queen’s contiguity or Euclidean distance calculated exponentially with \( \alpha = 2 \). Whereas \( D_0 \) is the Euclidean matrix performed with the average yield rate of the 113 regencies/ municipalities.
(iii) As opposed to the previous, in this step we will check whether the volatility of yield rate over the years should be included in this analysis. The $D_1$ in clustering is obtained from step (ii), while $D_0$ will arranged by: a. the average of yield rate only; and b. the average and volatility of yield rate. We conduct clusters of five in order to correspond the number of provinces in this analysis.

(iv) Selection of the zone for area yield, whether it is in form of province-based or new cluster-based, is evaluated by calculating the ratio of standard deviations within clusters ($S_W$) and standard deviations between cluster ($S_B$). The determination of the new clusters number at this stage is based on the number of dendrogram vertical lines which are being intersected by certain threshold line. The smaller value of ratio of $S_W$ and $S_B$ obtained, the better clustering algorithm is [8].

(v) If the new cluster-based is selected as the best zone, Kruskal-Wallis test and Dunn test will be conducted. It is used to ensure the simple adjustment of regions division (number of clusters in each provinces) using Ward-like hierarchical clustering.

3. Results and discussions

3.1. Selection of spatial dependence

The Moran test is carried out to check the spatial correlation between regencies/ municipalities because of pattern between neighbors that indicates similar characteristics. At 10 per cent significance level, there is sufficient evidence to state that there are spatial autocorrelations between regions with significance values of 0.083 and 0.075. The spatial weights used in this test are weight based on Euclidean distance, which calculated exponentially with $\alpha = 2$, and based on Queen contiguity matrix respectively.

Besides bringing out a smaller p-value, the Queen contiguity matrix is also selected as spatial information in this study because it delivers a better performance of clustering. Ward-like hierarchical clustering produces the ratios $S_W/S_B$ of 10.899 and 0.575, considers those spatial weights in geographical proximity of dissimilarity $D_1$.

3.2. Do we need the volatility as a variable considered in clustering analysis?

After gaining the best spatial information statistically, we check whether the volatility of yield rate over ten years is necessarily needed in clustering analysis. Volatility is a measure of the return dispersion which provides a single number intended to quantify the uncertainty or risk exposure [9]. In this case, volatility indicates the return behavior of the yield rates and helps estimate the fluctuations that may happen from 2007 to 2018.

| Table 2 Ratio between $S_W$ and $S_B$ to consider volatility in the model |
|-----------------|-----|-----|-----|
| Scenario       | Number of clusters | $S_W$ | $S_B$ | $S_W/S_B$ |
| a              | 5   | 0.122 | 0.212 | 0.575     |
| b              | 5   | 0.129 | 0.113 | 1.134     |

Scenario (a) and (b) on Table 1 are still conducted by Ward-like hierarchical clustering explained in the third steps of Section 2.3. The five clusters were taken in order to match the number of provinces involved in this study. Because the ratios $S_W/S_B$ of scenario (a) is smaller, the volatility will not be used in this further regencies/ municipalities cluster analysis. Nevertheless, the existence of volatility in clustering regions in other provinces should be reviewed. It is because the paddy productivity characteristics from year to year can vary.
3.3. Identification of zone-based area yield

The two zones to be compared are province-based and the new cluster-based as we have stated in the introduction. In this study, the new cluster-based, which used scenario (a), is identified as the best zone that can divide heterogeneous regions into smaller groups. Lower disaggregate level-based results the smaller ratio of \( S_W/S_B \), 0.711, than province-based, 9.346. The number of clusters within each province is determined by considering the dendogram produced by the Ward-like hierarchical method (Table 3).

**Table 3 Statistics summary of Kruskal-Wallis test and Dunn test**

| Scenario     | Number of clusters | p-value of Kruskal-Wallis test | Cluster comparison | p-value of Dunn test |
|--------------|--------------------|--------------------------------|-------------------|---------------------|
| West Java    | 3                  | 0.000                          | 1 vs 2            | 0.000               |
|              |                    |                                 | 1 vs 3            | 0.031               |
|              |                    |                                 | 2 vs 3            | 0.000               |
| Central Java | 3                  | 0.000                          | 1 vs 2            | 0.001               |
|              |                    |                                 | 1 vs 3            | 0.000               |
|              |                    |                                 | 2 vs 3            | 0.001               |
| DI Yogyakarta| 2                  | 0.157 *                        | 1 vs 2            | 0.079 *             |
| Eas Java     | 3                  | 0.000                          | 1 vs 2            | 0.000               |
|              |                    |                                 | 1 vs 3            | 0.002               |
|              |                    |                                 | 2 vs 3            | 0.000               |
| Banten       | 3                  | 0.062 *                        | 1 vs 2            | 0.228 *             |
|              |                    |                                 | 1 vs 3            | 0.013               |
|              |                    |                                 | 2 vs 3            | 0.048               |

P-value of Kruskal-Wallis and Dunn test with an (*) denotes differences among cluster within the province and pair of clusters means differ are not sufficient evidence respectively (\( \alpha = 5\% \)).

Due to that simple adjustment of regions division (Table 3), we had not been able to ensure that results of the cluster can be used immediately as the new clusters-based. Kruskal-Wallis test, which is a rank-based nonparametric test can be used to determine if there are statistically differences between clusters of each provinces [10], is conducted. On top of that, we implement Dunn test to perform multiple pairwise comparison [11]. Hence the pair of clusters that are not statistically different will be gathered.

By using a 95% confidence level, there is sufficient evidence to state that the average of yield rate between clusters in West Java, Central Java, and East Java Province are different (Table 3). So that the number of clusters in each of these provinces is three. Whereas in Banten and DI Yogyakarta Province, there are no significant differences. The Dunn test on Table 3 shows that pair of clusters, first and second, of these two provinces will be merged due to the insufficient evidence of average yield rate differences. Accordingly, the DI Yogyakarta becomes a cluster and Banten has two clusters.
Table 4 Characteristics of each cluster

| Cluster | n  | Mean | Median | Volatility | Level | LCL  | UCL  |
|---------|----|------|--------|------------|-------|------|------|
| JTG3    | 8  | 4.975| 5.082  | 0.119      | Low   | -    | -    |
| BNT1    | 6  | 5.343| 5.373  | 0.050      |       |      |      |
| BNT2    | 2  | 5.206| 5.206  | 0.066      |       |      |      |
| JBR3    | 10 | 5.480| 5.495  | 0.070      | Middle| 0.036| 0.133|
| JTG2    | 18 | 5.599| 5.606  | 0.089      |       |      |      |
| JTM2    | 17 | 5.782| 5.749  | 0.124      |       |      |      |
| JTM3    | 9  | 5.252| 5.313  | 0.107      |       |      |      |
| DIY     | 5  | 5.922| 6.095  | 0.074      |       |      |      |
| JBR1    | 4  | 6.069| 6.077  | 0.052      |       |      |      |
| JBR2    | 13 | 5.879| 5.856  | 0.058      | High  | 0.029| 0.111|
| JTG1    | 9  | 5.997| 5.926  | 0.077      |       |      |      |
| JTM1    | 12 | 6.206| 6.167  | 0.090      |       |      |      |

JTG is Central Java, JTM is East Java, BNT is Banten, JBR is West Java, DIY is DI Yogyakarta. Mean and Median are the central measure of average yield rate per cluster, and the unit of measurement is ton per ha. LCL and UCL refers to Lower and Upper of Control Limit chart respectively for the return volatility of yield rate in each label.

Table 4 explains the statistical summary of each cluster, the average yield rate, which is generalized into three simpler cover levels. The covers of zone are arranged according to the national target of paddy productivity provided by MoA, 5.46 tons per ha [12]. The margin of errors at 5 per cent level of 0.342 was obtained from the standard deviation of yield rate between 34 provinces in Indonesia at 2016. Clusters that have median value below 5.118 tons per ha are included in Low level, while High level for whom are above 5.802 tons per ha contrarily.

Its covers, namely Low, Middle, and High, formed to make the characteristics of the cluster is easier to be recognized nationally.

In order to investigate the stability of yield rate over the years which is being a new potential in designing a zone area yield in Java, the control limit is obtained (Table 4). Rather than the average, the return volatility of yield rates is used to represent risk characteristics corresponding to the in-control state. Bakshi and Madan [13] said that volatility rate is often associated with big swings, the higher return volatility will lead to the higher risk profile. It means the critical yield index \((y_c)\) of the cluster has a potential to shift instantaneously. As long as the cluster has a return volatility within the control limit such in Table 4, upper (UCL) and lower (LCL), the risk is assumed to be in control [14].

Now let us consider our new proposed cluster-based area yield scheme policy. The varied agricultural characteristics and paddy productivity as the basis risk of heterogeneity in implementing crop insurance can be resolved by developing cluster zone. In this design, the indemnity is linked to the reference index, \((y_c)\) and is the same for all farmers within the cluster. If on average, the realized productivity of a district or village drops under the \((y_c)\), then the indemnity will be triggered and all farmers within that district or village compensated subject to the insured farming land size.

4. Conclusion
Zone based area yield scheme in crop insurance is proposed to be an alternative policy to improve crop insurance in Indonesia, MPCI. Based on historical data, we have found that the
use of province based for the zone area yield scheme is not ideal. The new cluster-based, that is lower disaggregate of province, results the smaller ratio $S_W/S_B$ than province-based. It does not only utilize the average of paddy productivity, which becomes a new potential to obtain the critical yield index of the zone, but also the Queen contiguity matrix by regency/ municipality to cluster similar characteristics of farmers. The areas formed are three clusters on each West Java, Central Java, and East Java, two clusters on Banten, and DI Yogyakarta as a whole group. The return volatility of yield rate each cluster itself remains the risks are restrained.

Selecting the area or zone boundaries for an area yield using productivity has also never done before. This zone used to estimate the trigger yield that should be robust and reduce the influence of outliers. Because of the homogeneous and same loss distribution of risks, it can be assumed that the cluster expected value is the appropriate premium for the farmer on that cluster. Further identification of critical yield for each cluster should be done analytically.

References
[1] Pasaribu S M and Sudiyanto A 2016 Agricultural Risk Management: Lesson Learned from the Application of Rice Crop Insurance in Indonesia Climate Change Policies and Challenges in Indonesia. (Tokyo: Springer) pp 305-22
[2] Sutomo V A, Kusumaningrum, D, Anisa R and Paramita A 2019 A Bootstrap Simulation for Comparison of Group Risk Plan and Multi-Peril Crop Insurance Policy J. Phys.: Conf. Ser. 1366 012075
[3] Skees J R, Black J R and Barnett B J 1997 Designing and Rating an Area Yield Crop Insurance Contract American Journal of Agricultural Economics vol 79 (Oxford University Press on behalf of the Agricultural Applied Economics Association) pp 430-38
[4] Kusumaningrum D, Anisa R, Sutomo V A and Tan K S Alternative Area Yield Index Based Crop Policies in Indonesia (A Presented Paper of 9th International Conference on Mathematical and Statistical Methods for Actuarial Science and Finance (MAF 2020))
[5] Wang HH 2000 Zone-Based Group Risk Insurance Journal of Agricultural and Resource Economics 25 pp 1-21
[6] Chavent M, Kuentz-Simonet V, Labenne A and Saracco J 2018 ClustGeo: An R Package for Hierarchical Clustering with Spatial Constraints Computational Statistics vol 33 (Springer Verlag, In press) pp 1-24 (doi.org/10.1007/s00180-018-0791-1)
[7] Johnson R A, Wichern D W 2007 Applied Multivariate Statistical Analysis, Sixth Edition (New Jersey: Pearson Education, Inc) pp 680-96
[8] Bunkers M J, Miller J R and DeGaetano A T 1996 Definition of climate regions in the northern plains using an objective cluster modification technique J. Climate 9 pp 130-46
[9] Klugman S A, Panjer H H and Willmot G E 2012 Loss Models: From Data to Decisions, Fourth Edition (New Jersey: John Wiley Sons, Inc) pp 19-46
[10] Sprent P and Smeeton N C Applied nonparametric statistical methods, Third Edition (Florida: Chapmann Hall/CRC texts in statistical science series) p 206
[11] Dinno A 2015 Nonparametric pairwise multiple comparisons in independent groups using Dunn’s test The Stat Journal 15 pp 292-300
[12] Data and Information Center of Ministry of Agriculture 2016 Outlook Komoditas Pertanian (Jakarta: Data and Information Center of Ministry of Agriculture)
[13] Bhaksi G and Madan D 2006 A Theory of Volatility Spreads Management Science vol 52 (INFORMS) pp 1945-56
[14] Montgomery D C 2009 Introduction to Statistical Quality Control, Sixth Edition (New Jersey: John Wiley Sons, Inc) pp 180-90