A comparative analysis of clustering algorithms to identify the homogeneous rainfall gauge stations of Bangladesh

Sangita Paul*
University of Dhaka, Dhaka, Bangladesh - spaul@isrt.ac.bd

Mohammad Samsul Alam
University of Dhaka, Dhaka, Bangladesh - msalam@isrt.ac.bd

Abstract

Bangladesh has an agro-based economy which mainly depends on rain. Dealing with the rainfall at each of the stations separately is time consuming as well as subject to more errors. It seems more advantageous and flexible to deal with a group of homogeneous stations rather than individual stations. Different regions of Bangladesh experience different precipitations every year where some regions receive very much similar precipitations. To identify the homogeneous stations clustering algorithms of multivariate techniques are applied in this study and a comparison between these algorithms is made. Annual and seasonal (pre monsoon, monsoon and post monsoon) precipitation data of 30 stations from 1977 to 2012 recorded by Bangladesh Meteorological Department are used in this study. Fuzzy C-means, agglomerative hierarchical and K-means clustering methods are applied to classify the precipitation series and identify the hydrologically homogeneous groups. The optimal numbers of clusters are four, three, three and four chosen for annual, pre monsoon, monsoon and post monsoon rainfalls respectively using Gap statistic. The spatial distributions of stations in the clusters identified by each of the clustering methods are obtained for annual precipitations and for precipitations during pre monsoon, monsoon and post monsoon. The regional homogeneity test based on L-moments showed that the clusters identified by Fuzzy C-means method are sufficiently homogeneous compared to that by hard clustering methods, hierarchical and K-means. Thus it is recommended to prefer Fuzzy C-means method to classify the precipitation series and for identifying hydrologically homogeneous regions to those hard clustering methods.

Keywords: gap statistic; hierarchical; fuzzy c-means; L-moments.

1. Introduction

Bangladesh has a tropical monsoon-type climate, with a hot and a rainy summer and a dry winter. Rainy season lasts for more than two months with the exception of nearly three months winter, the summer lasts rest of the year. Most rain occur during the monsoon (June-September) and little in winter (November-February). Different regions of Bangladesh receive different precipitations every year where some regions receive very much similar precipitations. It is troublesome to deal with each and every rainfall regions rather than dealing with groups of similar rainfall regions for any analysis. Reliable decisions can be made by studying these groups only. Moreover, time and cost will be less (Bailey & Ken, 1994). Clustering allows to group these hydrologically homogeneous regions into different groups (Bailey & Ken, 1994).

Cluster analysis or clustering is the process of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. The groups which contain similar objects are called clusters. It is an exploratory data mining process, and a common technique for statistical data analysis, used in many fields. This study is designed to group hydrologically homogeneous regions of Bangladesh into some groups using agglomerative hierarchical, K-means and Fuzzy C-means (FCM) clustering methods. Dikbas et al. (2012) accomplished research on FCM clustering method and found regions defined by FCM are sufficiently homogeneous. Cluster analysis methods such as hierarchical and non-hierarchical clustering algorithms have been widely used in determining homogeneous regions. These are hard clustering methods where each feature vector either belongs to a certain cluster or not. There is a soft clustering method called Fuzzy C-means (FCM) where each feature vector can belong

to all clusters with a degree of membership between 0 and 1.

This study uses the secondary monthly rainfall data of 35 stations in Bangladesh collected by Bangladesh Meteorological Department (BMD) for several years till 2012. The data of 36 years from 1977 to 2012 of 30 stations are extracted for this study. Among these 5 of the stations, Tangail, Sayedpur, Chuadanga, Mongla and Ambagan are excluded because rainfall data of few years are available for these stations. It is said in the literature that the stations to be used in the identification of homogeneous regions should have statistically significant data ($n > 30$ years) (Dikbas et al., 2012) .

2. Methodology

In this study the optimal numbers of clusters are chosen using Gap statistic and then the stations are clustered using Fuzzy C-means method, Agglomerative hierarchical and K-means separately. The homogeneity test of clusters identified by these clustering methods is done based on L-moments ratio.

2.1. Gap statistic

A major challenge in cluster analysis is estimation of optimal number of clusters. Tibshirani et al. (2000) proposed a method (the gap statistic) for estimating the number of clusters (groups) in a data set. The technique uses the output of any clustering algorithm, comparing the change in within-cluster dispersion with that expected under an appropriate reference null distribution. The data set $\{x_{ij}\}$, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$ consists of p features measured on n independent observations. Let $d_{ii'}$ denote the distance between observations i and i'. $d_{ii'}$ is the squared Euclidean distance such that $\sum_j (x_{ij} - x_{i'j})^2$. The steps to find the Gap statistic are: i) Suppose we have clustered the data into k clusters $C_1, C_2, ...C_k$ with C_r denoting the index of observations in cluster r, and $n_r = |C_r|$, where "| |" represents the modulus.

$$D_r = \sum_{i,i' \in C_r} d_{ii'}, \tag{1}$$

be the sum of pair wise distances for all points in cluster r and

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r, \tag{2}$$

is the pooled within-cluster sum of squares around the cluster means. ii) Each reference feature is generated by sampling uniformly over the range of the observed values for that feature. Here p reference features are generated. Monte Carlo sample of size n are generated from each of the p reference features. Thus we get a $n \times p$ matrix of reference data set. Similarly B copies of reference data sets are generated. Each reference data set is clustered giving within dispersion measures, W^*_{kb}, $b = 1, 2, \ldots, B$ and $k = 1, 2, \ldots, k$. The expectation under the reference distribution is

$$E^*_n\{log(W_k)\} = (1/B) \sum_b log(W^*_{kb}). \tag{3}$$

The gap statistic is defined as

$$Gap_n(k) = E^*_n\{log(W_k)\} - log(W_k). \tag{4}$$

iii) let $\bar{l} = (1/B) \sum_b log(W^*_{kb})$. Accounting for the simulation error in $E^*_n\{log(W_k)\}$ results in the quantity

$$s_k = sd_k \sqrt{1 + 1/B}, \tag{5}$$

where sd_k is the standard deviation given by

$$sd_k = [(1/B) \sum_b \{log(W^*_{kb}) - \bar{l}\}^2]^{1/2}. \tag{6}$$

iv) A graph is plotted of $Gap_n(k)$ against k. The optimal number of clusters is then the value of k for which $Gap_n(k)$ has increased significantly with smallest s_k (Dudoit & Fridlyand, 2002). The reference data set's clusters have the most heterogeneity (have largest within-dispersion). At this value of k, $log(W_k)$ falls significantly below that of reference data set. Hence there will be k clusters of observed values those will have smallest within-dispersion.

2.2. Fuzzy C-means method

Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" but "fuzzy" in the same sense as fuzzy logic. Fuzzy logic is a form of many-valued logic; it deals with reasoning that is approximate rather than fixed and exact. Compared to traditional binary sets (where variables may take on true or false values), fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. The fuzzy cluster method was proposed by Dunn (1974) based on the fuzzy logic method and was developed and extended by Bezdek (1981). The most used fuzzy cluster method is the Fuzzy C-means method (FCM) which was proposed by Bezdek in 1981.The term "fuzzy logic" (Hazek, 2010) was introduced with the 1965 proposal of fuzzy set theory by Zadeh (1965). Fuzzy sets are sets whose elements have degrees of membership (possibility) to be in the set whose value range from 0 to 1.

Let x_k denote the k^{th} feature vector in the n dimensional dataset, $x_k = (x_{k1}, x_{k2}, \ldots, x_{kn})^T$, $x_k \in R^n$, n

is the number of variables and T denotes the transposition of a vector or matrix. Let N be the number of feature vectors. The equation (7) gives the matrix $\mathbf{X}$.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{pmatrix}. \tag{7}$$

The objective function to be minimized in the FCM method is given in equation (8). The restrictions given in equations (10) and (11) should be considered in the minimization of the objective function.

$$J(U, V : X) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^m d_{ik}^2(x_k, v_i), \tag{8}$$

where $i = 1, 2, \ldots, c$ and $k = 1, 2, \ldots, N$. c is the number of clusters and v_i is the i^{th} cluster center. $m \in [1, \infty]$ is the fuzziness weight term controlling the membership shared within the fuzzy clusters. A large m results in smaller memberships u_{ik} and hence, fuzzier clusters. In the limit $m = 1$, the memberships u_{ik} converge to 0 or 1, which implies a crisp partitioning. In the absence of experimentation or domain knowledge, m is commonly set to 2. The distance measurement between the k^{th} feature vector and the i^{th} cluster center is calculated using equation (9).

$$d_{ik}^2 = (x_k - v_i)^T A (x_k - v_i), \tag{9}$$

where A is the sample covariance matrix. In case of Euclidean distance there will be no sample covariance matrix, A in equation (9). Each feature vector belongs to a cluster with a degree of membership grade. As given in equation (10) the sum of the membership degrees of a feature vector is equal to 1.

$$\sum_{i=1}^{c} u_{ik} = 1 \qquad \forall k \in \{1, ...N\}, \tag{10}$$

$$0 < \sum_{k=1}^{N} u_{ik} < N \qquad \forall i \in \{1, ...c\}. \tag{11}$$

The membership matrix, $\mathbf{U}$ given in equation (12) shows the membership degrees of the feature vectors.

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1c} \\ u_{21} & u_{22} & \cdots & u_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N1} & u_{N2} & \cdots & u_{Nc} \end{pmatrix}, \tag{12}$$

where $u_{ik} \in [0, 1]$ denotes the membership degree of the k^{th} feature vector of i^{th} cluster. In fuzzy cluster analysis, the membership degrees or the cluster centers are, respectively, calculated at each iteration step according to equations (13) and (14) (Rao & Srinivas, 2006).

$$u_{ik} = \frac{\frac{1}{(d^2(x_k, v_i))^{1/(m-1)}}}{\sum_{i=1}^{c} \frac{1}{(d^2(x_k, v_i))^{1/(m-1)}}} \qquad \begin{cases} 1 \le i \le c \\ 1 \le k \le N \end{cases}, \tag{13}$$

$$v_i = \frac{\sum_{k=1}^{N} (u_{ik})^m x_k}{\sum_{k=1}^{N} (u_{ik})^m}. \tag{14}$$

2.3. Homogeneity test

Regional homogeneity test based on L-moments proposed by Hosking & Wallis (1993, 1997) is to be applied to test the homogeneity of clusters identified by different clustering methods in this study. A homogeneity test (H test) based on L-moment ratios (L-Cv, L-Cs, and L-Ck) was proposed for testing the homogeneity of the groups identified by cluster analysis. Suppose there are N feature vectors with $i = 1, 2, \ldots, N$. For i^{th} feature vector L-coefficient of variation (L-CV) is $t^{(i)}$, the coefficient of L-skewness (L-Cs) is $t_3^{(i)}$ and the coefficient of L-kurtosis (L-Ck) is $t_4^{(i)}$. The regional average L-moment ratios are defined as $\overline{L-Cv} = \sum_{i=1}^{N} n_i t^{(i)} / \sum_{i=1}^{N} n_i$, $\overline{L-Cs} = \sum_{i=1}^{N} n_i t_3^{(i)} / \sum_{i=1}^{N} n_i$ and $\overline{L-Ck} = \sum_{i=1}^{N} n_i t_4^{(i)} / \sum_{i=1}^{N} n_i$. The weighted standard deviation (V_1) for a region may be determined by using the equation (15) based on L-Cv ratios.

$$V_1 = \frac{\sum_{i=1}^{N} n_i (L - Cv^{(i)} - \overline{L-Cv})^2}{\sum_{i=1}^{N} n_i}. \tag{15}$$

In the equation (15), n_i is the number of data of the feature vector i; $L - Cv^i$ is the L-moment ratio of feature vector i. For evaluating the homogeneity measure, Kappa distribution with four parameters is fitted to the regional dataset. The parameters are $\overline{L-mean}, \overline{L-scale}, \overline{L-Cs}$ and $\overline{L-Ck}$ calculated from original regional dataset. Kappa distribution is used to obtain G homogeneous data regions and simulation is performed. Variations are calculated for each generated region and the standard deviation (σ_{V_1}) and the average (μ_{V_1}) of these variations are determined. Then, H1 measures were calculated according to Equation (16) for testing the homogeneity of the regions identified by cluster analysis.

$$H1 = \frac{V_1 - \mu_{V_1}}{\sigma_{V_1}}, \tag{16}$$

Table 1: The gap statistics with their associated standard errors

	annual		pre monsoon		monsoon		post monsoon	
k	$GAP_n(k)$	s_k	$GAP_n(k)$	s_k	$GAP_n(k)$	s_k	$GAP_n(k)$	s_k
1	0.181	0.053	0.373	0.055	0.242	0.055	0.170	0.033
2	0.232	0.036	0.221	0.032	0.302	0.036	0.183	0.026
3	0.220	0.029	**0.241**	**0.026**	**0.308**	**0.029**	0.189	0.027
4	**0.279**	**0.030**	0.244	0.025	0.307	0.029	**0.202**	**0.028**
5	0.283	0.033	0.258	0.027	0.297	0.034	0.198	0.027
6	0.264	0.036	0.233	0.028	0.281	0.035	0.184	0.029
7	0.242	0.036	0.244	0.032	0.268	0.034	0.190	0.031
8	0.263	0.038	0.236	0.034	0.251	0.038	0.176	0.030
9	0.257	0.040	0.227	0.034	0.283	0.041	0.180	0.033
10	0.293	0.042	0.213	0.037	0.315	0.045	0.184	0.037
11	0.230	0.039	0.198	0.035	0.230	0.037	0.129	0.036
12	0.235	0.040	0.199	0.039	0.294	0.040	0.143	0.042
13	0.277	0.041	0.201	0.043	0.298	0.041	0.106	0.040
14	0.286	0.047	0.209	0.046	0.291	0.041	0.113	0.046

where (σ_{V_1}) is the standard deviation of the values obtained from the simulation and (μ_{V_1}) is the average of these values.

The following criteria are proposed by Hosking & Wallis (1993) for interpreting the H1 values and for determining the homogeneity of the regions: If H1 $<$ 1 then the region is 'acceptably homogeneous', if $1 \leq$ H1 < 2 then the region is 'possibly homogeneous' and if H1 ≥ 2 then the region is 'definitely heterogeneous'.

3. Results and discussion

The optimal numbers of clusters are chosen using Gap statistic and then the stations are clustered using Fuzzy C-means method, Agglomerative hierarchical and K-means separately. The homogeneity test of clusters identified by these clustering methods is done based on L-moments ratio.

3.1. Optimal number of clusters

The optimal number of clusters is chosen by Gap statistic in this study. The minimum and maximum number of clusters for this study are set 2 and 14 $(\frac{30}{2} - 1)$ respectively.

From the Table 1 for annual rainfalls the significant increases in Gap statistics occur at $k = 2, 4, 8, 10, 13$ and 14 among them at $k = 4$ the standard error is the lowest which is 0.030. At this number of clusters the variation of stations within clusters falls farthest below the variation of simulated values within clusters. Thus 4 distinct clusters can be found. So the number of clusters is 4 for annual precipitations according to this method.

Similarly by analyzing the results of this method the number of clusters, 4, 3, 3 and 4 are chosen for annual, pre monsoon , monsoon and post monsoon precipitations respectively.

3.2. Fuzzy clustering method

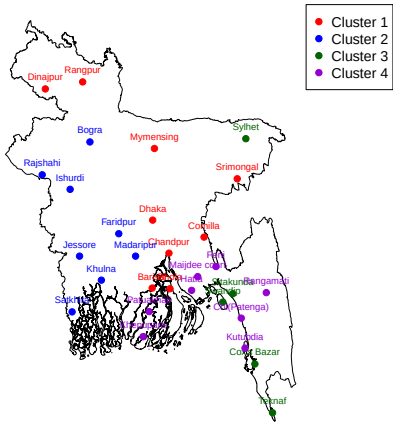
After applying FCM method we get the final membership degrees of stations to each cluster for annual, pre monsoon, monsoon and post monsoon precipitations. The first priority is to consider the stations with highest membership degrees: the station with the highest membership degree to a cluster belongs to that cluster. Cluster belongings (spatial distributions of stations in the clusters) are shown in Figure 1(a). From this figure northern stations - Rangpur and Dinajpur, all eastern stations except Sylhet and southern stations - Barisal and Bhola are in cluster 1, north-western stations and western stations are in cluster 2, few southern stations and south-eastern stations with Sylhet belong to cluster 3 and other southern and south eastern stations are in cluster 4. Similarly for pre monsoon, monsoon and post monsoon precipitations the spatial distributions of stations in the clusters based on the first priority are shown in Figures 1(b), 1(c) and 1(d) respectively.

Agglomerative Hierarchical and K-means clustering methods are also applied and clusters are found.

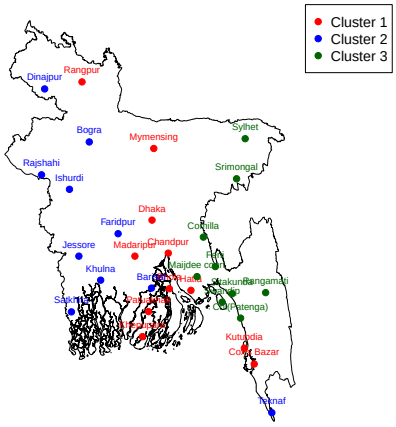
3.3. Regional homogeneity test

The regional homogeneity test results were evaluated from the Table 2 using the criteria proposed by Hosking & Wallis (1993) and findings are given below: For clusters defined by FCM clusters 1 and 2 are acceptably homogeneous (homo) and clusters 3 and 4 are possibly homogeneous for annual precipitations. All the clusters are acceptably homo for pre monsoon precipitations. Cluster 1 is possibly homo and clusters 2 and 3 are acceptably homo for monsoon precipitations. All are acceptably homo for post monsoon precipitations.

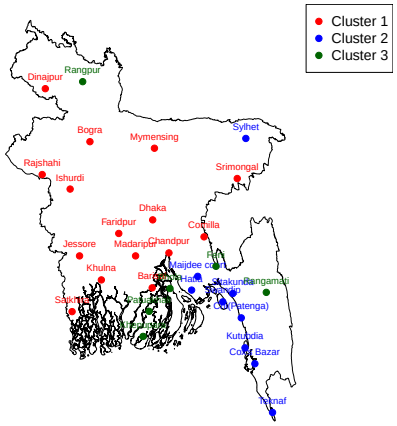
For clusters defined by hierarchical clustering method clusters 1,2 and 4 are acceptably homo and cluster 3 is definitely heterogeneous for annual. All the clusters are acceptably homo for pre monsoon. Clusters 1 and 3 are acceptably homo whereas cluster 2 is possibly homo for monsoon. Clusters 1 and 4 are possibly homo and others are acceptably homo for post mon-



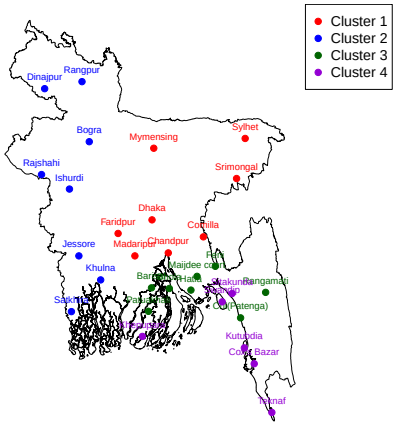
(a) for annual preccipitations



(b) for pre monsoon precipitations



(c) for monsoon precipitations



(d) for post monsoon precipitations

Figure 1: Clusters formed by fuzzy method based on the first priority

soon.

For clusters defined by K-means method clusters 1, 2 and 4 are acceptably homo but cluster 3 is definitely heterogeneous for annual. Clusters 1 and 3 are acceptably homo and cluster 2 is possibly homo for pre monsoon. Clusters 1 and 3 are acceptably homo but cluster 2 is possibly homo for monsoon. Cluster 1 is possibly homo but clusters 2,3 and 4 are acceptably homo for post monsoon.

4. Conclusions

In this study 30 stations of Bangladesh are used to classify the hydrologically homogeneous regions by Fuzzy C-means, hierarchical and K-means methods. The monthly precipitation data of 30 stations taken from Bangladesh Meteorological Department is used for this study. The hydrological homogeneous regions are formed for annual precipitations and also for precipitations during pre monsoon, monsoon and post monsoon. The optimal number of clusters are determined using Gap statistic. It was best to choose 4

Table 2: Regional homogeneity test for clusters

		Clustering method					
		FCM (first priority)		Hierarchical		K-means	
Precipitations	Cluster no.	no. of stations	value of H1	no. of stations	value of H1	no. of stations	value of H1
annual	1	9	0.65[a]	17	0.52[a]	7	0.23[a]
	2	8	-0.03[a]	3	-0.10[a]	11	0.33[a]
	3	5	1.39[b]	8	2.18[c]	8	2.50[c]
	4	8	1.89[b]	2	-0.07[a]	4	0.12[a]
pre monsoon	1	11	0.20[a]	12	0.09[a]	16	0.19[a]
	2	10	0.12[a]	17	0.05[a]	12	1.63[b]
	3	9	0.83[a]	1	0[a]	2	-0.16[a]
monsoon	1	15	1.31[b]	17	0.56[a]	18	0.51[a]
	2	9	0.90[a]	11	1.66[b]	8	1.82[b]
	3	6	0.45[a]	2	-0.28[a]	4	-0.53[a]
post monsoon	1	8	-0.76[a]	13	-1.25[b]	6	-1.70[b]
	2	8	0.54[a]	8	0.64[a]	5	0.68[a]
	3	8	0.05[a]	4	-0.87[a]	7	0.11[a]
	4	6	0.30[a]	5	1.45[b]	12	-0.11[a]

[a]acceptably homogeneous
[b]possibly homogeneous
[c]definitely heterogeneous

clusters for annual precipitations, 3 clusters for pre monsoon precipitations, 3 clusters for monsoon precipitations and 4 clusters for post monsoon precipitations so that the within sum of squares of clusters are minimum. The regional homogeneity of the regions identified by different clustering methods are tested using regional homogeneity test (H test) based on L-moments method. H1 values for all clusters defined by FCM method based on the first priority are lower than the limit value, 2. For annual precipitations 1 cluster defined by hierarchical method and 1 cluster defined by K-means method contain heterogeneity and clusters defined by these methods are also less homogeneous than the clusters defined by FCM method.

In comparison of FCM method with other hard clustering methods such as agglomerative hierarchical method or K-means method it is seen that homogeneity test results of clusters identified by FCM method are better than that of hard clustering methods. Thus FCM method classify the precipitation series more accurately than the others. Therefore, spatial distribution of stations in the clusters identified by FCM is recommended as the best.

References

Bailey, & Ken. (1994). Numerical taxonomy and cluster analysis. Typologies and Taxonomies, page 34.

Bezdek, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Dikbas, F., Firat, M., Koc, A. C., & Gungor, M. (2012). Classification of precipitation series using fuzzy cluster method. International journal of climatology, 32: 1596 - 1603.

Dudoit, S., & Fridlyand, J. (2002). A prediction-based re-sampling method for estimating the number of clusters in a database. Genome Biology, 3(7).

Hajek, P. (2010). Fuzzy logic. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Fall 2010 edition.

Hosking, J. R. M., & Wallis, J. R. (1993). Some statistics useful in regional frequency analysis. Journal of Climate, 29(2): 271 - 281.

Hosking, J. R. M., & Wallis, J. R. (1997). Regional Frequency Analysis: An Approach Based on L - moments. Cambridge University Press, Cambridge, UK.

Tibshirani, R., Walther, G., & Hastie, T. (2000). Estimating the number of clusters in a data set via the gap statistic. Journal of Royal Statistical Society, 63(2): 411 - 423.

Zadeh, L. (1965). Fuzzy sets. Information and control, 8: 338 - 353.