Spatial -Temporal Climate Regionalization Using Clustering Methods Over Taiwan

Manikandan Sathianarayanan1, Pai-Hui Hsu\*2

1,2Department of Civil Engineering, National Taiwan University, Taiwan

Abstract

Over the past few decades, there has been a substantial shift in the Earth's climate, leading to the emergence of global warming and its detrimental impacts on both human life and various activities. Within this framework, identifying climate zones holds significant importance due to its potential to prevent or effectively address unforeseen natural disasters. Many current efforts focused on this objective are constrained to individually studying distinct climate variables associated with a given climate zone. In this paper, we have outlined several hierarchical clustering algorithms, including K-means and Agglomerative Hierarchical clustering methods and applied it to four remote sensing data variables: precipitation, temperature, elevation, and vegetation. All the algorithms were applied on a pixel-by-pixel basis instead of utilizing the locations of weather stations, which has the potential to accurately identify various climate zones across Taiwan. The strong similarity between the identified climate zones and the Köppen-Geiger climate classification map validates the precision and effectiveness of the proposed method.

Keywords – Climate zone, K-means, Agglomerative Hierarchical clustering

1. Introduction

Discretization methods play a crucial role in attaining a comprehensible grasp of intricate geospatial datasets. An illustrative instance is provided by global climate categorizations, such as the Koppen-Geiger approach. The process of global climate classification involves categorizing regions with similar climatic conditions, depicting average climatic patterns in different areas by analysing certain key climate variables. Discretizing numerous local climates (LCs) into a limited set of climate types simplifies the spatial climate variability for more meaningful and accessible analysis. Hence, climate classification offers an instinctive and invaluable comprehension of the connections between climate and the Earth's physical and biological mechanisms. These encompass erosion(Peel, Finlayson et al. 2007), soil characteristics(Rohli, Joyner et al. 2015), biological communities(Baker, Diaz et al. 2010), as well as the distribution of invasive species (Werier and Naczi 2012) and vectors for viruses. Climate classification aids in visualizing global climate datasets, enabling the depiction of climate change by portraying shifts in the geographic boundaries of key climate types. Likewise, it facilitates the visualization of projected spatial distributions of climate types based on climate models’ predictions.

This study explores a clustering methodology for Taiwan classifying climates. The clustering procedures organizes local climates in to clusters founded on their shared similarities, utilizing an automatic clustering algorithm. These clusters are then linked with climate types. The CTs are essentially unveiled by an algorithm, relying solely on the data's inherent attributes, devoid of any preconceived notions about their anticipated characteristics or geographical distribution. Leveraging comprehensive Taiwan’s Long Term Monthly Means climatic datasets(Hijmans, Cameron et al. 2005),a viable clustering approach emerges. Each dataset entry characterizes a local climate (LC) at a station or grid cell. A dissimilarity function gauges LC dissimilarity, used with LC representation to evaluate climate type (CT) uniformity, regardless of clustering or KGC origin. In this study, we employed distinct techniques for selecting and preparing input data which is similar to KG, and established a novel clustering approach that deviates from previous methods.

1. Dataset

This study utilized time series data from MODIS Terra vegetation index, captured at 1km x 1km spatial resolution with monthly coverage. Employing a constrained view maximum value compositing approach, the images aimed to mitigate cloud and atmospheric effects while adhering to specific viewing angles(Huete, Didan et al. 2002). Acquired from the Land Processes Distributed Active Archive Center (LP DAAC), a set of 964 images spanning February 2000 to December 2020 covered the study area. MODIS Reprojection Tool (MRT) facilitated band extraction, format conversions, and projection adjustments. NDVI image preprocessing encompassed conversion to GeoTIFF, sinusoidal-to-WGS84/Albers projection transformation, and Savitzky-Golay filtering (Jönsson and Eklundh 2004).

Additionally, Land Surface Temperature (LST) data derived from MODIS sensors on Terra and Aqua satellites, with 1km resolution, was studied. MOD11A1 day time daily LST products for 2000 to 2020 were acquired from the USGS website. LST computation relied on the split-window algorithm, incorporating thermal infrared bands 31 and 32, compensating for atmospheric effects, and referencing a global land surface emissivity look-up table(Snyder, Wan et al. 1998, Wan, Zhang et al. 2002).

The joint GPM initiative by NASA and JAXA, initiated in 2014, aims to offer global precipitation insights. GPM satellites feature key tools: GMI for assessing precipitation attributes and DPR for examining storm structures. These sensors set benchmarks for integrating rainfall data from other satellites in the constellation. IMERG products, particularly Level-3 options, find broad scientific and operational utility. These include Early, Late, and Final Runs, with the latter incorporating gauge calibration. Our study focuses on GPM, TRMM's successor, providing Level-3 (IMERG-Final) precipitation data at 0.1° resolution from 60°S to 60°N. These data are accessible at varying intervals via NASA's site (Huffman, Bolvin et al. 2015). IMERG daily precipitation is derived from 48-hour accumulation data, scaled by 0.5 for accurate precipitation estimation.

1. Study area

The research site is located in East Asia, situated on the western edge of the Pacific Ocean, encompassing the mainland of Taiwan along with a few distant small islands. To the east of the Taiwanese mainland lie the Pacific Ocean, Bashi Channel, Taiwan Strait, and East China Sea, in sequence. Encompassing an area of 36,197 square kilometers, the study region spans from 120°E to 122°E and from 22°N to 25°N. This region is primarily characterized by mountain ranges in the eastern portion and gradually sloping plains in the western part. Covering approximately 394 kilometers in length from north to south and 144 kilometers in width from west to east.

1. Methods

Cluster analysis divides data into smaller segments by grouping alike objects (such as climate stations or grid points) into appropriate categories, while distinguishing dissimilar ones. Clustering algorithms fall into two main categories: hierarchical and non-hierarchical. Both approaches use distance or correlation measures to evaluate similarity (short distance, strong correlation) or dissimilarity (long distance, weak correlation) among objects, such as individual grid points.

4.1 Principle component analysis (PCA)

Principal Components Analysis (PCA) stands as the prevailing multivariate statistical method utilized to diminish the dimensionality of complex multivariate datasets(Jolliffe 1987). By performing orthogonal rotations on the coordinate axes of the initial variables, a transformation is achieved that aligns them with prominent clusters of data points within the multivariate space. This process typically reveals a small number of new axes, referred to as principal components (PCs), which predominantly capture the variability present in the dataset. The interpretation of the physical or other processes that have worked on the data is based on the relationships between those variables that contribute significantly to the primary components that account for the largest amounts of the variability. We implemented Z-score normalization to standardize each of the data variables. The determination of the number of retained PCs for cluster analysis was achieved by examining the scree plot. Utilizing the location of the elbow in a scree plot to choose PCs is a recognized approach for concluding PCA(Jackson 1993). The depicted figure illustrates a scree plot that displays the percentage of variance explained by individual components as well as the cumulative variance. Notably, this scree plot specifically highlights the initial seven components, all of which collectively elucidate over 95% of the total variance. Nonetheless, the initial component on its own accounts for slightly less than 60% of the variance, indicating that additional components could be necessary for a comprehensive understanding. However, the cumulative effect of the first six principal components clarifies over 90% of the overall variability in the standardized ratings. This rationale validates the retention of the first six principal components as a reasonable approach for further dimensionality reduction. Furthermore, as observed from the figure, it becomes evident that the initial six components elucidate more than 95% of the dataset. Consequently, our subsequent analysis exclusively concentrates on these specific components.

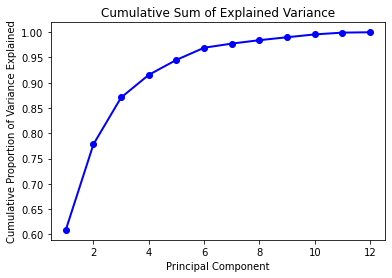
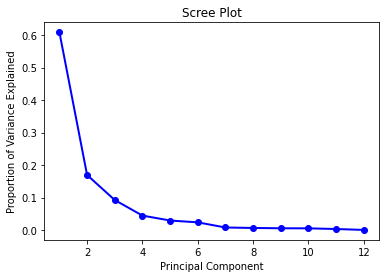


Figure 1. scree plot showing the proportion of variance explained against principle component and cumulative proportion of variance explained

4.2 Hierarchical clustering methods

In climate research, the predominant hierarchical clustering techniques are average linkage and Ward's methods. Ward's approach seeks to minimize the sum of squares within clusters, whereas the average linkage method aims to minimize within-cluster variance while maximizing between-cluster variance. This is achieved by blending characteristics of the single linkage and complete linkage methods(Kalkstein, Tan et al. 1987). Due to its tendency to favor clusters with similar member counts, Ward's method might lead to an instance being assigned to a more distant cluster rather than a closer one, especially if the distant cluster has significantly fewer members. Using variance instead of sum of squares in the average linkage method mitigates the impact of cluster sizes on clustering dependency. This alteration enhances the realism of groupings achieved through the average linkage method, resulting in more accurate and meaningful clusters. Ward’s (Ward Jr 1963) and average linkage (Murtagh 1983)methods have been most widely used in climate analyses(Kalkstein, Tan et al. 1987, Fovell and Fovell 1993) and the same was utilized in this study.

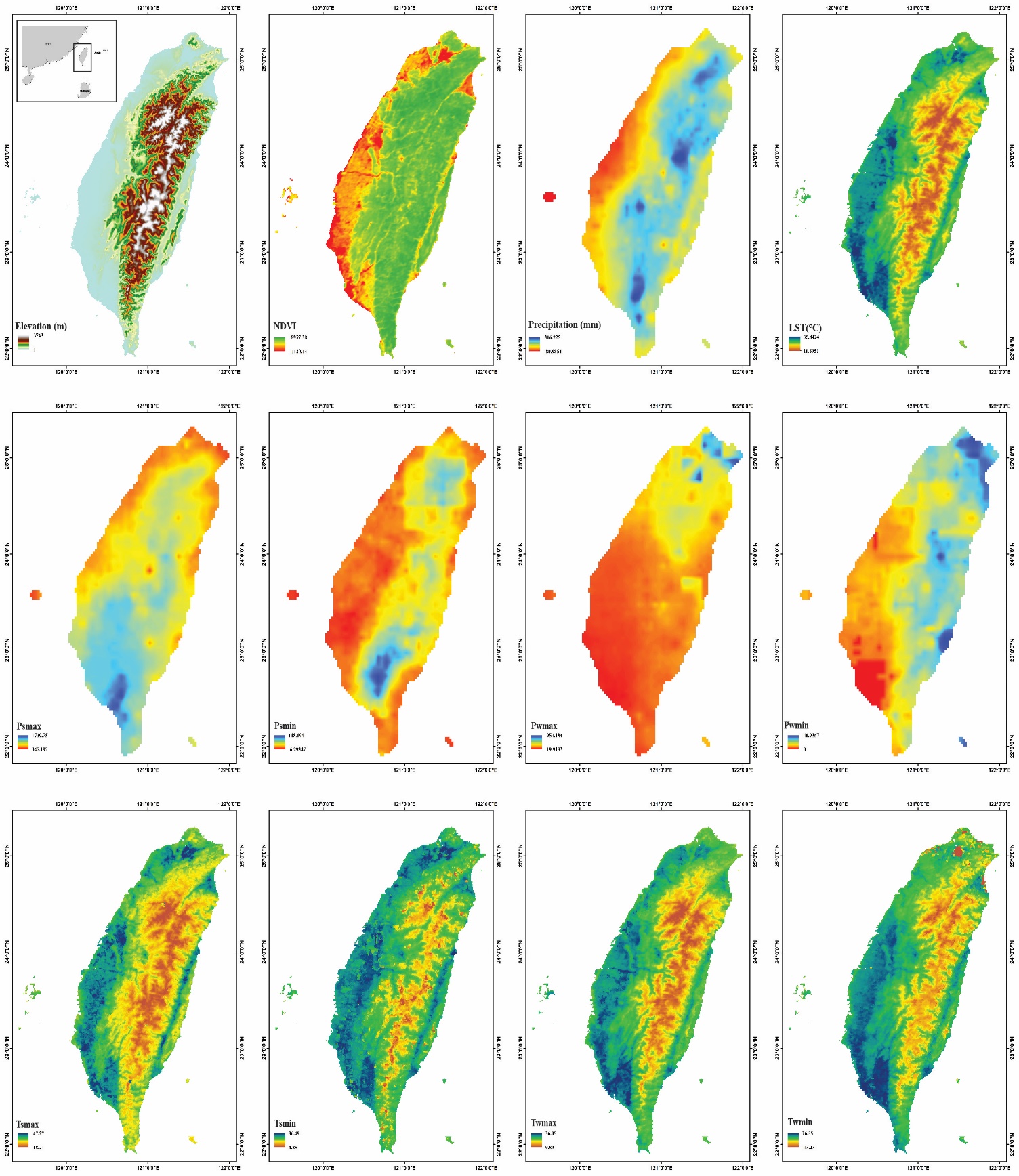
4.3 Non- Hierarchical clustering algorithms

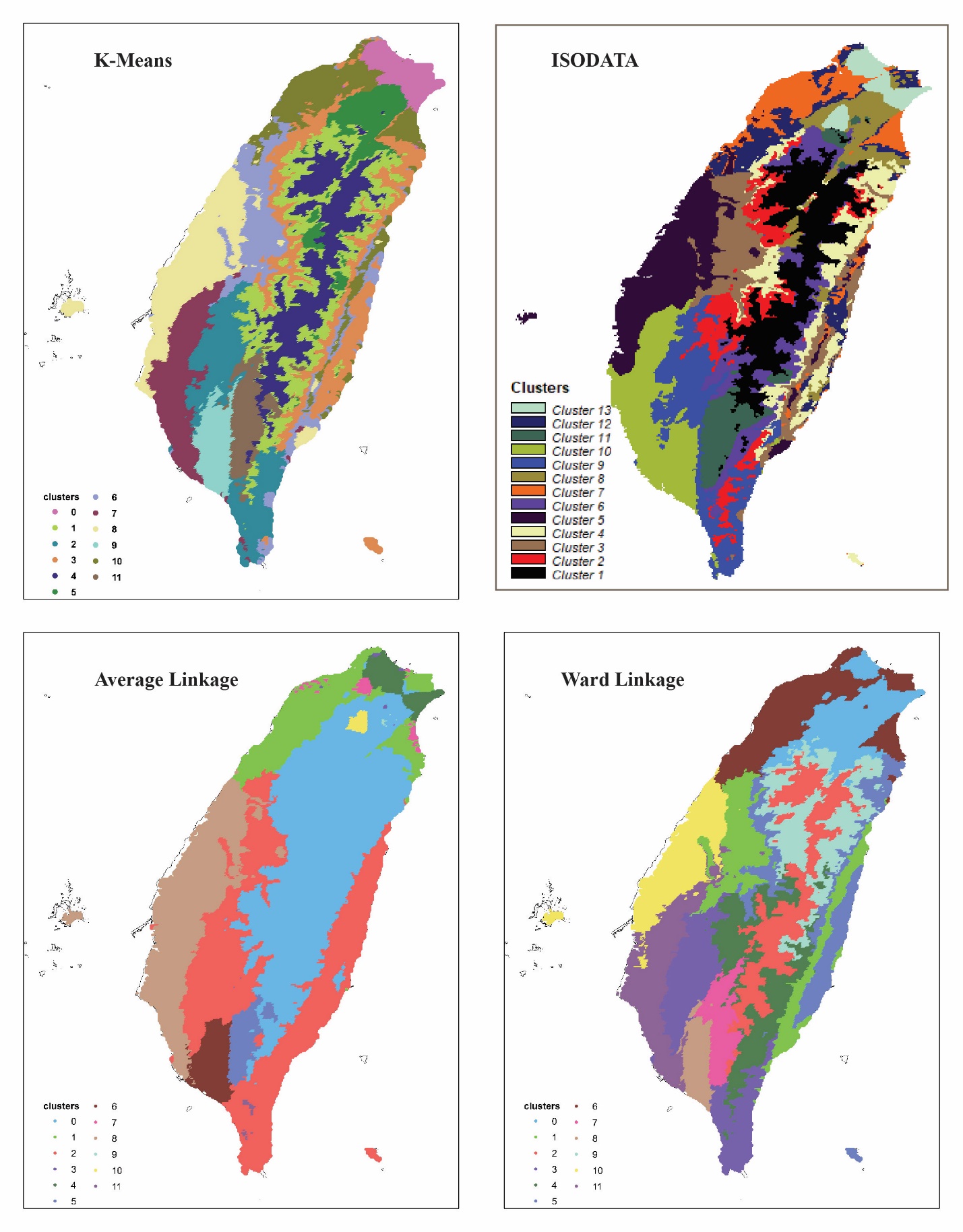
Hierarchical clustering algorithms initiate with individual clusters, each containing a single member, and they prohibit the reassignment of each observation. Conversely, non-hierarchical clustering algorithms possess the capability to reassign each observation to a distinct cluster based on proximity to the nearest centroid. The nonhierarchical methods are intended to group data entities into a unified classification consisting of 'k' clusters, where the value of 'k' is either predetermined or established during the clustering process. A significant limitation of nonhierarchical cluster analysis is the requirement for pre-existing knowledge about the quantity of clusters(Carvalho, Melo-Gonçalves et al. 2016), or in our context, the number of climate regions. Though,(Gong and Richman 1995) was observed that non-hierarchical approaches exhibited superior performance compared to hierarchical methods when evaluated using growing season precipitation data in the central and eastern regions of the United States. For this investigation, K-means was employed to conduct nonhierarchical clustering of climate regions. The clustering count was established based on Köppen climate classifications. We employed the K-means method to categorize the grid cells around Taiwan into a specific number of climate types. Ultimately, the climate types were classified using the K-means method.

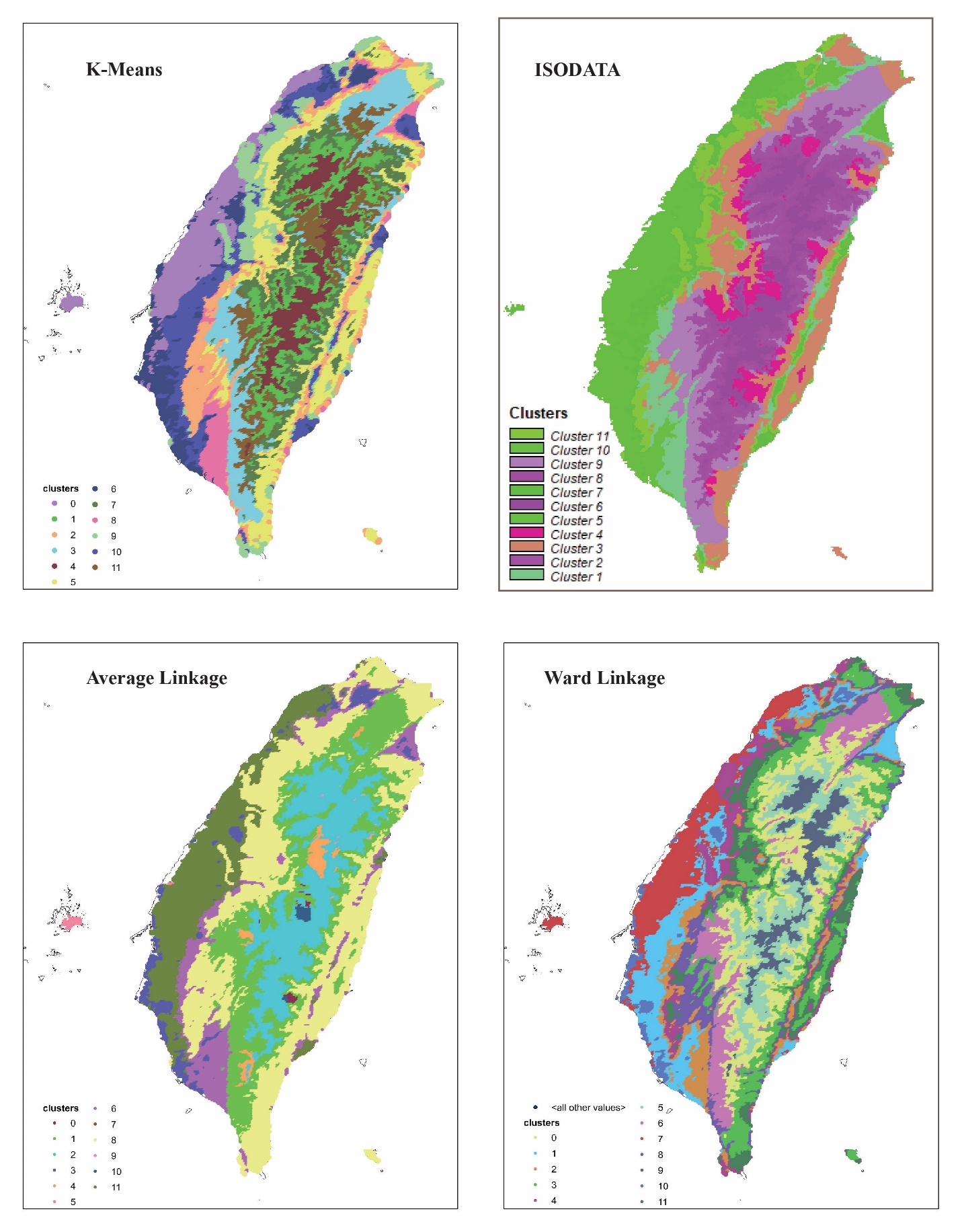
4.4 ISODATA clustering

The iterative self-organizing data analysis technique (ISODATA) is a variant of the k-means clustering algorithm and belongs to the category of partitioning-type clustering methods. This approach finds broad application in domains like image analysis, including fields such as remote sensing and medical sciences(Bachmann, Donato et al. 2002, Pan, Li et al. 2003).In contrast to k-means clustering, ISODATA clustering eliminates the need for predefining the number of classes. Initially set with a user-defined cluster count, the ISODATA algorithm adjusts the cluster count through merging, splitting, or removing clusters, guided by specific heuristics. This iterative process aims to converge towards a solution featuring an optimal number of clusters. Within this research, ISODATA clustering was executed utilizing the efficient implementation of the ISODATA algorithm available within the SAGA GIS package.

1. Result and discussion

Figure 2. Input variables used in this study along with study area

Figure 3. Clustering of 12 input variables using various clustering methods.

Figure 4. clustering with four input variables

The long-term datasets for precipitation, temperature, and vegetation were further categorized into specific components: mean annual temperature (Tmean), mean annual precipitation (Pyear), air temperature of the coldest month in summer (Tsmin), air temperature of the warmest month in summer (Tsmax), air temperature of the coldest month in winter (Twmin), air temperature of the warmest month in winter (Twmax), precipitation of the driest month in summer (Psdry), precipitation of the wettest month in summer (Pswet), precipitation of the driest month in winter (Pwdry), and precipitation of the wettest month in winter (Pwwet). We employed two sets of input data for climate classification preparation: one comprising 12 variables including all(case-1), and the other containing only four variables (Precipitation, Temperature, Elevation, and Vegetation-case-2). The optimal number of clusters determined is equivalent to the Köppen-Geiger (KG) zones of Taiwan, totaling 11. Figure 3 displays the clustering outcomes of K-means, ISODATA, Average Linkage, and Ward Linkage methods, utilizing all 12 input variables. These methods yielded 11 clusters each, except for ISODATA, which generated 13 clusters. This discrepancy is due to the fact that ISODATA clustering doesn't necessitate prior specifications of class numbers. Figure 4 displays the results of different clustering methods with only four input variables.

Case-1: Relatively smaller climate clusters were identified consistently in the northern portion of Taiwan across all clustering methods, while common clusters emerged in the central region. Compared to other methods, the average linkage clustering approach tends to produce more generalized large spatial clusters rather than smaller ones.

Case-2: Small and more overlap climate clusters were identified in the northern and eastern side of Taiwan for all methods except ISODATA. Results from all four methods show that as elevation increases, an increasing number of closely matched climate clusters are identified in the central region.

1. Concluding remarks and Future works

Our study aimed to investigate naturally occurring clusters within climate data over Taiwan and intended to do juxtapose the resultant climatic regions with those obtained through the KG classification system. We objectively outlined global climatic regions by conducting cluster analyses on a data matrix encompassing 10 climatic variables and elevation. Our results show that there are number of small climate clusters exist in all clustering methods, alongside a shared climate cluster in the central region. Additional research is warranted to conduct a comparative analysis between the results of climate clusters and the KG climate classification zones. This investigation aims to ascertain the climate similarities between the clustering outcomes and the KG classification.

References:

Bachmann, C. M., T. F. Donato, G. M. Lamela, W. J. Rhea, M. H. Bettenhausen, R. A. Fusina, K. R. Du Bois, J. H. Porter and B. R. Truitt (2002). "Automatic classification of land cover on Smith Island, VA, using HyMAP imagery." IEEE Transactions on Geoscience and Remote Sensing **40**(10): 2313-2330.

Baker, B., H. Diaz, W. Hargrove and F. Hoffman (2010). "Use of the Köppen–Trewartha climate classification to evaluate climatic refugia in statistically derived ecoregions for the People’s Republic of China." Climatic Change **98**: 113-131.

Carvalho, M. J., P. Melo-Gonçalves, J. C. Teixeira and A. Rocha (2016). "Regionalization of Europe based on a K-Means Cluster Analysis of the climate change of temperatures and precipitation." Physics and Chemistry of the Earth, Parts A/B/C **94**: 22-28.

Fovell, R. G. and M.-Y. C. Fovell (1993). "Climate zones of the conterminous United States defined using cluster analysis." Journal of climate **6**(11): 2103-2135.

Gong, X. and M. B. Richman (1995). "On the application of cluster analysis to growing season precipitation data in North America east of the Rockies." Journal of climate **8**(4): 897-931.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones and A. Jarvis (2005). "Very high resolution interpolated climate surfaces for global land areas." Int. J. Climatol. **25**: 1965-1978.

Huete, A., K. Didan, T. Miura, E. P. Rodriguez, X. Gao and L. G. Ferreira (2002). "Overview of the radiometric and biophysical performance of the MODIS vegetation indices." Remote sensing of environment **83**(1-2): 195-213.

Huffman, G. J., D. T. Bolvin, E. J. Nelkin and J. Tan (2015). "Integrated Multi-satellitE Retrievals for GPM (IMERG) technical documentation." Nasa/Gsfc Code **612**(47): 2019.

Jackson, D. A. (1993). "Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches." Ecology **74**(8): 2204-2214.

Jolliffe, I. T. (1987). "Rotation of principal components: some comments." Journal of Climatology **7**(5): 507-510.

Jönsson, P. and L. Eklundh (2004). "TIMESAT—a program for analyzing time-series of satellite sensor data." Computers & Geosciences **30**(8): 833-845.

Kalkstein, L. S., G. Tan and J. A. Skindlov (1987). "An evaluation of three clustering procedures for use in synoptic climatological classification." Journal of Applied Meteorology and Climatology **26**(6): 717-730.

Murtagh, F. (1983). "A survey of recent advances in hierarchical clustering algorithms." The computer journal **26**(4): 354-359.

Pan, Y., X. Li, P. Gong, C. He, P. Shi and R. Pu (2003). "An integrative classification of vegetation in China based on NOAA AVHRR and vegetation-climate indices of the Holdridge life zone." International Journal of Remote Sensing **24**(5): 1009-1027.

Peel, M. C., B. L. Finlayson and T. A. McMahon (2007). "Updated world map of the Köppen-Geiger climate classification." Hydrology and earth system sciences **11**(5): 1633-1644.

Rohli, R. V., T. A. Joyner, S. J. Reynolds and T. J. Ballinger (2015). "Overlap of global Köppen–Geiger climates, biomes, and soil orders." Phys. Geogr. **36**: 158-175.

Snyder, W. C., Z. Wan, Y. Zhang and Y.-Z. Feng (1998). "Classification-based emissivity for land surface temperature measurement from space." International Journal of Remote Sensing **19**(14): 2753-2774.

Wan, Z., Y. Zhang, Q. Zhang and Z.-l. Li (2002). "Validation of the land-surface temperature products retrieved from Terra Moderate Resolution Imaging Spectroradiometer data." Remote sensing of Environment **83**(1-2): 163-180.

Ward Jr, J. H. (1963). "Hierarchical grouping to optimize an objective function." Journal of the American statistical association **58**(301): 236-244.

Werier, D. A. and R. F. C. Naczi (2012). "Carex secalina (Cyperaceae), an introduced sedge new to North America." Rhodora **114**: 349-365.