

USO DE ESTADÍSTICOS DESCRIPTIVOS PARA INTERPRETAR LOS RESULTADOS DE UN ANÁLISIS CLÚSTER; CASO PARTICULAR.

Use of Descriptive Statistics for Interpreting the Results of a Cluster Analysis; Particular Case

RESUMEN

El presente artículo muestra cómo el uso de estadísticos descriptivos de medidas de localización, se convierten en una herramienta auxiliar de la técnica multivariada análisis clúster en la interpretación de las fases y comportamientos de variables Macro-Climáticas que presentan correlación con patrones de temperaturas. El procedimiento se realiza a un conjunto de índices Macro-Climáticos y a estaciones con series de temperaturas en el departamento del Chocó. La metodología anterior permitió describir y explicar el comportamiento de los patrones de temperatura dadas las variaciones de transito de una fase a otra de las variables Macro- Climáticas utilizadas en este estudio.

PALABRAS CLAVES: Análisis clúster, Estadísticos descriptivos, Variables Macro-Climáticas, patrones de temperatura.

ABSTRACT

This article shows how the use of descriptive statistics of measures of location, become an auxiliary tool of cluster analysis multivariate technique in the interpretation of the phases and behavior of macro-climatic variables that have correlation with temperature patterns. The procedure is performed to a set of macro-climate indices and temperature series stations in the department of Chocó. The previous methodology was to describe and explain the behavior patterns of temperature variations due to the transit from one phase to another of the macro-climatic variables used in this study

KEYWORDS: Cluster analysis, descriptive statistics, macro-climatic variables, temperature patterns.

1. INTRODUCCIÓN

El análisis clúster o de conglomerados, es una de las técnicas multivariadas de clasificación automática o no supervisada y de reconocimientos de patrones sin supervisión, que tiene por objetivo agrupar elementos en grupos homogéneos en función de las similitudes o similitudes entre ellos. Sin embargo en muchos temas esta técnica no es suficiente para lograr una descripción de las cualidades y características de las variables objetos de estudios; por ello en algunos casos se debe recurrir a técnicas de la estadística descriptiva uní variada como herramienta auxiliar en la construcción de las interpretaciones que se derivan de los resultados, tal es el caso de este estudio.

2. DESARROLLO TEÓRICO

Fecha Recepción: 9 de Septiembre de 2010

Fecha aceptación: 15 de Noviembre de 2010

JORGE ANDRÈS URRUTIA

M. Sc Investigación Operativa y Estadística.

Docente Auxiliar

Universidad Tecnológica de Pereira

Jurutia@utp.edu.co

REINER PALOMINO

M. Sc Ciencias Meteorología

Docente Investigador

Universidad Tecnológica del Chocó

reiner@utch.edu.co

IVAN RENE GALINDO E

Ingeniero Industrial

Profesor Auxiliar

Universidad Antonio Nariño - Roldanillo

email: ivan.galindo@uan.edu.co

Semillero de Investigación INDUANROL (UAN - Roldanillo)

2.1 ANÁLISIS CLÚSTER DE K-MEDIAS.^[1]

La interpretación de los niveles de temperatura producto de las incidencias de las variables Macro-Climáticas, se realiza mediante la aplicación del análisis Clúster de K-Medias que es un método no jerárquico de agrupación o de clasificación. Es de anotar que los métodos No-Jerárquicos, están diseñados para la clasificación de individuos (no de variables) en K grupos. El procedimiento es elegir una partición de los individuos en K grupos e intercambiar los miembros de los Clúster para tener una partición mejor, metodología propuesta por (Pérez Cesar 2006). La medida de asociación puede

ser una distancia o una similaridad. Cuando se elige una distancia como medida de asociación (por ejemplo la distancia euclídea) los grupos formados contendrán individuos parecidos de forma que la distancia entre ellos

ha de ser pequeña. Cuando se elige una medida de similitud (por ejemplo el coeficiente de correlación) los grupos formados contendrán individuos con una similitud alta entre ellos.

Las medidas de asociación entre individuo, está dada por distancia χ^2

$$\chi^2 = m \left[\sum_{i=1}^p \sum_{j=1}^q \frac{n_{ij}^2}{m_i \cdot m_j} - 1 \right] \quad (1)$$

2.2 ESTADÍSTICOS DESCRIPTIVOS DE LOCALIZACIÓN (CUARTILES)^[2]

Los cuartiles son los tres valores de la variable que dividen a un conjunto de datos ordenados en cuatro partes iguales, Q_1 , Q_2 , Q_3 que determinan valores correspondientes a 25%, 50% y 75% e las observaciones; Q_2 coincide con la mediana. Cuando se tiene una matriz de datos la expresión matemática que determina el cálculo es

$$Q_k = L_i + \frac{\left[\frac{K \cdot N}{4} \right] - F_{i-1}}{f_i} * a_i \quad (2) \quad K = 1, 2, 3$$

Donde

L_i es el límite inferior de la clase donde se encuentra el cuartil.

N es la suma de las frecuencias absolutas.

F_{i-1} es la **frecuencia acumulada** anterior a la clase del cuartil.

a_i es la amplitud de la clase.

2.3 VARIABLES MACRO-CLIMÁTICAS

Se entiende como variables Macro-Climáticas al conjunto de variables Oceánicas y atmosféricas de carácter multidecadal con influencia en las condiciones meteorológicas del planeta. Dichas variables presentan comportamientos de tendencia que se denominan fases; las variables utilizadas este caso son: PDO, NAO, ONI, MEI, NIÑO1+2, NIÑO3+4 Y NIÑO 4. Los dos primeros índices presentan dos fases positiva y negativa, mientras que los otros índices presentan tres fases, positiva negativa y normal.

2.4. TEMPERATURA^[3].

La temperatura se entiende como una medida de energía. La temperatura está relacionada con la energía calorífica de los rayos solares y es importante porque determina la formación de las nubes, afecta los valores de humedad atmosférica o cantidad de vapor de agua que se encuentra en el aire, e influye en la presión atmosférica, es decir, la fuerza que ejerce el peso del aire sobre la superficie terrestre su unidad más común es °C.

3. DESARROLLO

El análisis clúster se aplica posterior a la realización de un análisis de componentes principales y a un análisis de correlación canónica. El análisis de componentes principales se le aplicó por separado tanto a las variables Macro -Climáticas como a las estaciones con patrones de temperatura; el objetivo era establecer estadísticamente grupos de índices con comportamientos similares y grupos de estaciones con tendencia similares en las series de temperaturas. Posteriormente se aplicó un análisis de correlación canónica para establecer qué tipo de correlación presentaban las variables Macroclimáticas con las series de temperatura y para luego poder realizar un análisis Clúster de K-medias para evaluar los efectos que ocasionan las variaciones de dichas variables en los patrones de temperatura.

3.1 MATRIZ DE DATOS FORMADA PARA EL ESTUDIO.

La Tabla 1 presenta un resumen de los datos de las variables Macro-Climática; la tabla 2, muestra los datos estandarizados de las estaciones con patrones de temperatura. En la construcción de estas tablas, se verificó en cumplimiento del supuesto de normalidad.

AÑO	MES	NAO	PDO	ONI	MEI	NIÑO1+2	NIÑO 3	NIÑO3.4	NIÑO4
1883-1990	EN	13.89	7.53	4.3	7.608	3.26	1.6	586.09	4.15
1883-1990	FB	13.7	9.56	4.2	6.64	4.65	3.41	591.02	2.7
1883-1990	MZ	-7.29	16.71	4.3	9.821	7.54	4.49	600.9	1.92
1883-1990	AB	11.16	15.85	4	12.173	8.04	3.41	612.41	1.95
1883-1990	MY	-5.85	18.89	3.8	14.934	1.8	0.77	613.19	2.63
1883-1990	JN	-12.26	12.16	4	12.935	-1.82	0.47	607.55	3.87
1883-1990	JL	-1.97	16.99	4.6	10.129	-3.05	0.67	599.08	5.41
1883-1990	AG	29.71	11.82	4.5	9.004	-2.93	0.4	590.14	5.56
1883-1990	SP	3.48	5.85	3.7	6.835	-0.88	0.18	588.5	6.02
1883-1990	OT	-16.75	1.39	2.5	6.362	0.14	-0.01	586.91	5.97
1883-1990	NV	6.25	0.61	1.9	5.003	0.04	-0.26	584.83	6.16
1883-1990	DC	15.41	3.54	1.3	4.557	0.02	-0.79	582.83	4.52

TABLA 1. MATRIZ DATOS DE LOS ÍNDICES MACRO-CLIMÁTICOS.

LORO	SAN ISIDRO	APTO EL CA	TERESITA L	SAUTATA	SAN JOSE P	APTO CONDO	ISTMINA	PANAMERICA
1.288	1.955	1.533	0.998	-0.860	1.641	2.034	1.865	-0.795
-1.069	0.055	1.073	0.224	0.146	1.920	2.301	2.083	-0.276
1.497	1.996	1.185	0.464	0.349	2.232	2.258	2.034	-0.004
1.298	2.213	0.596	1.345	2.221	2.165	-0.009	1.858	1.572
1.943	2.084	0.770	0.511	1.667	2.160	2.380	1.813	1.269
1.701	1.469	0.765	0.977	1.496	2.117	1.840	1.695	0.890
1.497	1.315	0.547	1.159	1.744	2.164	0.335	1.694	0.785
1.147	1.259	-0.236	1.029	1.469	1.866	0.086	1.292	0.324
0.421	0.326	-1.189	0.592	0.382	0.196	0.734	0.381	-0.274
-0.322	0.864	-0.753	-0.531	-0.540	-0.348	0.230	0.622	-1.062
0.858	0.799	-0.144	0.712	0.921	0.138	0.716	0.635	-1.713
0.065	0.492	-1.245	0.309	-0.641	-0.890	-0.603	-0.455	-2.073
-0.374	-0.551	-0.396	-1.439	-0.026	-0.854	-0.926	-0.318	0.066
-0.860	-1.935	-0.767	-1.087	-0.581	-0.991	-0.997	-0.849	-1.134
-0.386	-0.257	-0.202	-0.704	0.349	-0.250	0.002	0.028	-0.214
-1.016	-0.487	-0.246	-1.168	-0.694	-0.221	-0.317	-0.124	-0.321

TABLA 2. MATRIZ DATOS DE SERIES ESTANDARIZADAS DE TEMPERATURA.

4. RESULTADOS Y ANÁLISIS

Para la aplicación del análisis Clúster, para mirar la relación de influencia de los índices Macro-climáticos en los patrones de temperatura, nos basaremos en la interpretación que nos arrojó el Análisis de Componentes Principales realizado para los Índices, Es de recordar que el ACP, arrojó dos componentes, en las cuales se observó que el índice AMO, aportaba poca variabilidad, por debajo de 2%, razón por la cual fue excluida del análisis. Sin embargo se observó que al correr un nuevo ACP, la primera componente recogió los índices (ONI, MEI, NIÑO1+2, NIÑO3, NIÑO3.4 y NiÑO4), quienes presentan tendencia de comportamiento y variabilidad similares, definidas por dos fases, *Positiva* y *Negativa* y en la segunda Componente, los índices (PDO y NAO), de los cuales consideraremos tres fases, *Negativa*, *Normal* y *Positiva*.

Con el objetivo de entender estas interacciones, a partir del análisis Clúster, aplicando el algoritmo de K- medias para los índices de la primera componente con los patrones de temperatura, se usará el estadístico descriptivo de posición cuartíl, y se compararan con los valores de los clúster, bajo el criterio de que patrones de temperaturas por encima del tercer cuartíl, obedece a al estado de verano y por debajo del primero, corresponde a la fase del invierno. El proceso de comparar los patrones Meteorológicos con los índices (NAO y PDO), se utilizaran el primero, segundo y tercer cuartíl, con el objetivo de interpretar sus tres fases.

4.1 NANÁLISIS CLUSTER PARA LOS ÍNDICES (MEI, ONI, NIÑO 1+2, NIÑO 3, NIÑO 3.4 Y NIÑO 4) Vs PATRONES DE TEMPERATURA.

Para la interpretación de la relación entre índices y Patrones de temperatura, se hace necesario retomar los resultados de la correlación canónica, que mostró que estos tienen correlación positiva con la temperatura. Lo que quiere decir que para valores Altos de los índices en la fase positiva, mayor será la intensidad de la temperatura.

La interpretación de los Clúster se da a continuación:

	Conglomerado	
	1	2
LORO	-.5520964	.7673558
SANISIDRO	-.5977226	.8227417
APTOCARA	-.5704397	.7853086
TERESITA	-.5844319	.8286556
SAUTATA	-.4911069	.6154526
SANPABL	-.4876454	.6898453
APTOCONDO	-.5025700	.6894871
ISTMINA	-.6166426	.8565853
PANAMER	-.2746730	.3047061
ONI	-.62	.82
MEI	-.37	1.31
NIÑO12	-.65	.97
NIÑO3	-.78	.90
NIÑO3.4	26.22	27.86
NIÑO4	-.45	.67

TABLA 3 COMGLOMERADOS PARA ESTACIONES CON PATRONES DE TEMPERATURA E ÍNDICES MACROCLIMÁTICOS DE LA PRIMERA COMPONENTE

				VARIABLES	CLÚSTER1	CLÚSTER2
				ONI	-0.62142857	0.81538462
				MEI	-0.36951786	1.31053846
				NIÑO12	-0.645	0.9725641
				NIÑO3	-0.78107143	0.90102564
N DE TEMPE	T BAJAS	T NORMALES	T ALTAS	NIÑO3.4	26.2151786	27.8610256
CUARTILES	25	50	75	NIÑO4	-0.45339286	0.67282051
LORÓ	-0.775872133	-0.035515384	0.64546817	LORÓ	-0.5520964	0.76735585
SANISIDRO	-0.718419655	-0.109715703	0.52483163	SANISIDRO	-0.59772265	0.82274165
APTOCARA	-0.712382676	-0.144146629	0.59637324	APTOCARA	-0.57043967	0.78530862
TERESITA	-0.824715687	-0.182272422	0.73988001	TERESITA	-0.58443188	0.82865546
SAUTATA	-0.718189892	-1.35671E-14	0.58803889	SAUTATA	-0.4911069	0.61545264
SANPABL	-0.706976544	-0.23188479	0.40664492	SANPABL	-0.48764538	0.68984533
APTOCONDO	-0.603474727	-0.069309857	0.65595079	APTOCONDO	-0.50256997	0.68948707
ISTMINA	-0.702008612	0.009245742	0.63515065	ISTMINA	-0.61664261	0.85658528
PANAMER	-0.706450862	0.166848477	0.66530714	PANAMER	-0.27467305	0.30470608

TABLA 4 CLÚSTER Y CUARTILIS PARA ESTACIONES CON PATRONES DE TEMPERATURA E ÍNDICES MACROCLIMÁTICOS DE LA PRIMERA COMPONENTE

La tabla 4, muestra que para valores del primer cuartil que sean menores al Clúster uno se hablará de bajas temperaturas y los valores del tercer Cuartil que estén por encima del clúster dos, se hablaran de altas temperatura. El análisis es igual para los índices PDO Y NAO, pero en tres clúster, dados las tres fases de los índices.

	Conglomerado		
	1	2	3
LLORO	-.6358184	.1604758	1.0438888
SANISIDRO	-.5968779	.1908072	.9632812
APTOCARA	-.6153697	.0883551	1.0717878
TERESITA	-.5930562	-.0239088	1.1847067
SAUTATA	-.5179396	.0250314	.8183267
SANPABL	-.5717690	.5271198	.6924845
APTOCONDO	-.5585165	.4693610	.6404229
ISTMINA	-.6603467	.2463568	1.0481763
PANAMER	-.2024943	-.0772948	.2691516
NAO	-.15	-.78	2.41
PDO	.47	.62	1.61

TABLA 5 COMGLOMERADOS PARA ESTACIONES CON PATRONES DE TEMPERATURA E ÍNDICES MACROCLIMÁTICOS DE LA SEGUNDA COMPONENTE

N DE TEMPE	T BAJAS	T NORMALES	T ALTAS	VARIBLES	CLÚSTER1	CLÚSTER2	VARIBLES
				NAO	-0.15061224	-0.7832	2.4105
CUARTILES	25	50	75	PDO	0.47387755	0.62	1.608
LLORO	-0.77587213	-0.03551538	0.64546817	LLORO	-0.63581842	0.16047582	1.04388884
SANISIDRO	-0.71841966	-0.1097157	0.52483163	SANISIDRO	-0.59687786	0.19080721	0.96328124
APTOCARA	-0.71238268	-0.14414663	0.59637324	APTOCARA	-0.6153697	0.08835514	1.07178784
TERESITA	-0.82471569	-0.18227242	0.73988001	TERESITA	-0.59305615	-0.0239088	1.18470672
SAUTATA	-0.71818989	-1.3567E-14	0.58803889	SAUTATA	-0.5179396	0.02503145	0.8183267
SANPABL	-0.70697654	-0.23188479	0.40664492	SANPABL	-0.571769	0.52711976	0.69248449
APTOCONDO	-0.60347473	-0.06930986	0.65595079	APTOCONDO	-0.55851646	0.46936103	0.64042285
ISTMINA	-0.70200861	0.00924574	0.63515065	ISTMINA	-0.66034674	0.24635677	1.04817631
PANAMER	-0.70645086	0.16684848	0.66530714	PANAMER	-0.20249426	-0.07729485	0.26915159

TABLA 6 CLÚSTER Y CUARTILS PARA ESTACIONES CON PATRONES DE TEMPERATURA E ÍNDICES MACROCLIMÁTICOS DE LA PRIMERA COMPONENTE

De igual forma, se observa en la tabla 6 que para los índices PDO y NAO, la existencia de la relación, se presenta de la siguiente forma: para valores del cuartil uno menores que el clúster uno, las temperaturas serán muy bajas; para valores del cuartil dos que sean mayores que el clúster uno y menores que el clúster dos, serán normales, y las que están por encima del clúster tres en el tercer cuartil, serán altas. Estas relaciones se entenderán en sentido contrario para el índice NAO, dado que la correlación canónica de los patrones de temperatura con este índice es negativa.

5. CONCLUSIONES Y RECOMENDACIONES

Para este tipo de estudios el análisis clúster mostró ser una herramienta de baja ponencia en el tratamiento estadístico de datos climáticos, sin embargo en ayuda de estadísticos descriptivos, permitió establecer que durante el período de estudio correspondiente desde 1983 a 1990,

el 50% de los registros de precipitación, mostraron episodio de periodos con altos niveles de humedad y bajos niveles de temperatura, sobre todo en la región sur central del departamento.

De igual forma, se pudo establecer que los índices que mayor ponderación presentan en la explicación de los cambios de temperatura son los índices PDO; ONI; NIÑO1+2 y NIÑO 4, traduciéndose en que para altas temperatura superficial del mar y niveles bajos de presión, los niveles de temperatura serán altos.

El estudio permitió comprobar estadísticamente, que los valores máximos que alcanzó la temperatura en el departamento en el periodo de 1983 a 1990 fueron de 29 °C y valores mínimos de temperatura de mínimos de 19°C

El análisis clúster de K-medias subalterno con el estadístico descriptivo de medida de posición mostró que desde 1983 hasta el 2004, los índices ONI; MEI, NIÑO1+2, NIÑO3, NIÑO3.4 y NIÑO4, tomaron valores negativos en el 75% de las veces a lo largo de 22 años, lo que se traduce en que en este lapso de tiempo se presentó altos niveles de precipitación y bajos niveles de temperatura.

Se recomienda para futuros trabajos hacer solo uso de los índices ONI; MEI; NIÑO1+2, NIÑO 4, PDO, y NAO.

6. BIBLIOGRAFÍA

[1] Peña, D. 2002. Análisis de Datos Multivariantes. Madrid: McGraw Hills/Interamericana de España.

[2] Walpole, Myers. 2007 Probabilidad y Estadística para Ingeniería y Ciencia

[2] PÉREZ LÓPEZ, César, Técnicas de Análisis Multivariante de Datos, Aplicaciones con SPSS, Madrid, Universidad Complutense de Madrid."

[3] JORGE URRUTIA. Tesis de Maestría, Aplicación de Análisis Multivariado en la Determinación de la incidencia de variables Macroclimáticas en Patrones de Precipitación y Temperatura 2010.

[4] FRANCISCO SOLEY, Aplicación de Análisis Multivariado al Campo de Anomalías de la Precipitación en Centro America.