4.5 ANÁLISIS DE COMPONENTES PRINCIPALES APLICADA A LAS VARIABLES DE ESTUDIO

En esta sección, encontramos la solución de componentes principales utilizando los datos originales, estandarizados, y rotación ortogonal por el método de varimax (usando la matriz de correlación).

4.5.1. Determinación de las componentes principales usados los datos originales

El primer paso para la solución del análisis de componentes principales es determinar los valores propios (denotados como λ_1, λ_2, λ_3, $\cdots$, λ_p) de la matriz de varianza y covarianza ($\widehat{\Sigma}$), y la proporción de variación explicada por cada componente. Se puede probar que λ_i representa la varianza de la i-ésima componente principal.

TABLA XCVIII
Tabla de valores propios utilizando la matriz de covarianza

i	Valor propio (l_i)	% de variación para el i-ésimo componente	% Acumulado
1	**476413.47**	**98.87955**	**98.87955**
2	3446.295	0.71528	99.59483
3	1053.568	0.21867	99.81350
4	575.264	0.11940	99.93289
5	276.623	0.05741	99.99030
6	42.083	0.00873	99.99904
7	1.479	0.00031	99.99935
8	1.101	0.00023	99.99957
9	0.644	0.00013	99.99971
10	0.531	0.00011	99.99982
11	0.423	0.00009	99.99991
12	0.265	0.00006	99.99996
13	0.107	0.00002	99.99998
14	0.083	0.00002	100.00000

En la tabla XCVIII se muestran los valores característicos de cada componente principal, el porcentaje de variación explicada y variabilidad acumulado para la i-ésima componente, con la primera componente principal el porcentaje de explicación es del 98.88%.

TABLA XCIX
Estimación de las varianza para cada variable utilizando la matriz de varianza y covarianza

Variable	Varianza
X_1	2.594
X_2	26.036
X_3	0.012
X_4	0.017
X_5	0.029
X_6	0.081
X_7	0.000
X_8	0.012
X_9	0.001
X_{10}	63.020
X_{11}	1552.657
X_{12}	469019.991
X_{13}	4974.925
X_{14}	774.125

Para seguir desarrollando esta solución, debemos observar las varianza (varianza común) para las variables de la componente principal, ver tabla XCIX, las cuales indican el porcentaje de la i-ésima variable explicada por la solución propuesta, claramente observamos que la variabilidad está mal distribuidas en las variables, es decir en algunos casos las variables tienen varianzas superiores y otras son inferiores.

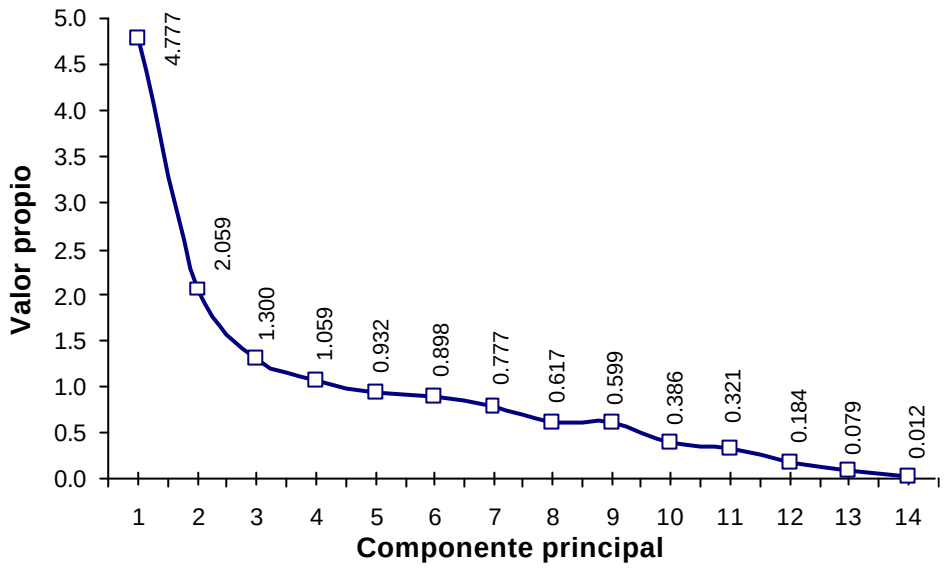
4.5.2. Determinación de las componentes principales usados los datos estandarizados.

La solución anterior de la sección 4.5.1, no es adecuada porque la distribución de la varianza común no es considerablemente uniforme, lo que deseamos encontrar son componentes principales que tengan una mejor explicación con las variables. Para esto utilizaremos los datos estandarizados, que es lo mismo utilizar la matriz de correlación. En la siguiente tabla, se muestra los valores propios, el porcentaje de la variación explicada y la variabilidad acumulada.

TABLA C
Tabla de valores propios utilizando la matriz de correlación

i	Valor propio (l_i)	% de variación para el i-ésimo componente	% Acumulado
1	**4.777**	**34.119**	**34.119**
2	**2.059**	**14.708**	**48.827**
3	**1.300**	**9.286**	**58.113**
4	**1.059**	**7.567**	**65.680**
5	**0.932**	**6.660**	**72.340**
6	0.898	6.415	78.755
7	0.777	5.548	84.303
8	0.617	4.406	88.709
9	0.599	4.276	92.985
10	0.386	2.757	95.742
11	0.321	2.293	98.035
12	0.184	1.314	99.349
13	0.079	0.562	99.911
14	0.012	0.089	100.000

GRÁFICO 4.1
Varianza explicada para cada componente principal utilizando la matriz de correlación



Fuente: Dirección nacional de estudios (Base de dato años lect. 1995–99)

Observando la tabla C, se ha escogido 5 componentes principales, cuyo porcentaje de explicación es del 72.34%, este criterio (porcentaje-explicación) es útil porque el investigador decide con cuanta información desea trabajar. Otros criterio es de seleccionar los valores propios mayores que uno, también coincide con los 5 factores seleccionados, esto claramente se pude observar en el anterior gráfico.

TABLA CI
Matriz de correlación entre la componente principal y la variable, utilizando la matriz de correlación lineal

	C_1	C_2	C_3	C_4	C_5	Varianza
X_1	**-0.469**	**0.795**	-0.003	0.240	-0.036	0.911
X_2	**-0.485**	**0.801**	0.010	0.236	-0.016	0.932
X_3	**-0.486**	**0.429**	0.177	-0.047	0.254	0.518
X_4	0.290	-0.305	**-0.469**	**0.483**	0.267	0.702
X_5	**-0.557**	-0.120	**-0.411**	0.183	-0.118	0.541
X_6	**0.441**	-0.054	0.216	-0.186	**0.626**	0.671
X_7	-0.039	-0.232	**0.675**	0.023	-0.359	0.640
X_8	-0.129	-0.386	0.265	**0.579**	-0.212	0.616
X_9	-0.094	-0.070	**0.513**	**0.448**	**0.394**	0.633
X_{10}	**0.796**	0.256	-0.014	0.033	-0.209	0.744
X_{11}	**0.890**	0.292	0.070	-0.001	-0.056	0.886
X_{12}	**0.879**	0.278	0.113	0.022	-0.048	0.865
X_{13}	**0.829**	0.224	-0.021	0.107	-0.099	0.759
X_{14}	**0.763**	0.072	-0.161	0.305	0.020	0.707
λ_j	4.777	2.059	1.300	1.059	0.932	
% explicación acumulada: 72.34						

Procediendo con el análisis de componentes principales, la tabla anterior muestra la matriz de correlación entre las variables de estudio y las componentes principales, los valores con negrillas corresponden a las cargas relativamente altas mayores que 0.40 (en valor absoluto). Analizando la matriz de carga (tabla CI), observamos que la primera componente principal estaría compuesta por 10 variables, mientras que es resto por menos de 4 variables, incluso repitiéndose entre componente, quitando de esta manera representatividad a las demás componentes principales.

4.5.3. Determinación de las componentes principales usados los datos estandarizados con rotación por el método de varimax

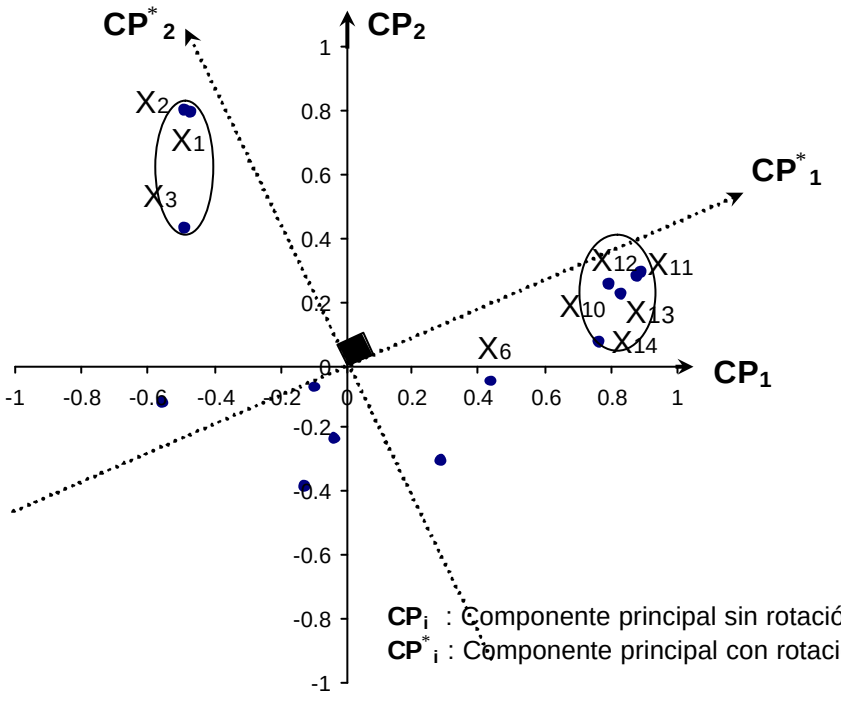
Aún no se puede mostrar la solución definitiva, ya que en algunos casos la i-ésima variable puede estar significativamente correlacionada con dos componente principal, para reducir esto utilizaremos el método de varimax, que consiste en retribuir la varianza de las variables dentro de ellas.

TABLA CII
Matriz de correlación entre la componente principal y la variable, utilizando la matriz de correlación con una rotación VARIMAX

	C_1	C_2	C_3	C_4	C_5	Varianza
X_1	-0.099	**0.978**	-0.071	-0.064	-0.026	0.976
X_2	-0.098	**0.969**	-0.074	-0.070	-0.031	0.961
X_3	-0.165	**0.265**	-0.139	-0.039	-0.043	0.121
X_4	0.051	-0.121	**0.973**	0.044	-0.059	0.970
X_5	**-0.316**	0.080	0.010	-0.142	-0.074	0.132
X_6	0.187	-0.114	0.046	**0.963**	0.014	0.978
X_7	-0.044	-0.047	-0.056	0.013	**0.989**	0.985
X_8	-0.098	-0.070	0.029	-0.028	0.079	0.022
X_9	-0.035	0.002	-0.015	0.006	0.056	0.005
X_{10}	**0.727**	-0.071	0.015	0.007	-0.033	0.535
X_{11}	**0.931**	-0.088	0.017	0.146	-0.028	0.897
X_{12}	**0.944**	-0.089	0.030	0.137	-0.009	0.920
X_{13}	**0.760**	-0.119	0.068	0.026	-0.040	0.599
X_{14}	**0.502**	-0.119	0.204	0.100	-0.041	0.319
λ_j	4.129	2.378	1.240	1.220	1.161	
% explicación acumulada:		72.34				

La tabla CII, muestra la matriz de correlación una vez rotada por el criterio de varimax, los valores con negrillas corresponden a las correlaciones relativamente altas (mayores que 0.40 en valor absoluto), podemos observar una más clara agrupación de las variables con sus componentes principales.

GRÁFICO 4.2
Guayas, colegios fiscales: Comparación gráfica de la correlación lineal de las dos componentes principales usando datos estandarizados con y sin rotación



***Fuente**:* Dirección nacional de estudios (Base de dato año lectivo 1997–98)

Observando los ejes CP_1 vs CP_2, las variables X_1, X_2, y X_3 están correlacionas entre sí, pero no podemos decir con cual componente

principal. Para esto se hace necesario una rotación ortogonal de las componentes principales (usando el criterio varimax), los nuevos ejes coordenados están representados por CP^*_1 vs CP^*_2, donde claramente se puede notar que variables están mas correlacionadas con la componente principal (ver gráfico 4.2).

Componente principal 1: Las variables más importantes son las siguientes:

- **X_5 :** Clasificación por sexo del estudiante del colegio
- **X_{10}:** Número de aulas en la institución educativa
- **X_{11}:** Personal del colegio
- **X_{12}:** Número de estudiantes promovidos
- **X_{13}:** Número de estudiantes no promovidos
- **X_{14}:** Número de estudiantes desertores

A esta componente principal se la podría denominar “Nivel de calidad del colegio fiscal”, por el alcance de las variables contenidas en él, y explica el 29.49% del total de la varianza de la población.

Componente principal 2: En esta componente prevalecen las siguientes variables:

- $\mathbf{X_1}$: Cantón al que pertenece la institución.
- $\mathbf{X_2}$**:** Parroquia a la que pertenece la institución.
- $\mathbf{X_3}$**:** Zona (urbana o rural) al que pertenece la institución

A esta componente principal lo denominamos "ubicación geográfica del colegio fiscal", y posee el 16.99% de explicación con respecto a la variación de la población.

Componente principal 3: Esta constituida mayoritariamente por una sola variable:

- $\mathbf{X_4}$: Tipo de jornada que tiene la institución.

La denominamos con el mismo nombre original por contener a una sola variable y posee 8.86% de explicación con respecto a la variación de la población.

Componente principal 4 y 5: En estas componentes principales prevalecen una sola variable, éstas al igual que el factor anterior se la denominaran con el mismo nombre, y poseen 8.71% y 8.29% de la variación total de la población respectivamente.

Con todo lo anterior, se puede llegar a la conclusión que es posible representar a cada unidad por medio de componente principales usando datos estandarizado con una rotación ortogonal por el criterio de varimax, en la que se puede observar una clara agrupación de las características en los factores que buscan el mayor porcentaje de explicación con respecto a la variación de la población. A continuación se presentan los vectores ortonormalizados que determinan las componentes, ver tabla CIII.

TABLA CIII
Tabla de valores propios con sus respectivos vectores propios de cada componente principal usando varimax

	Vectores Propios				
Variable	b_1	b_2	b_3	b_4	b_5
X_1	-0.215	**0.554**	-0.003	0.233	-0.037
X_2	-0.222	**0.558**	0.009	0.229	-0.017
X_3	-0.222	**0.299**	0.155	-0.046	0.263
X_4	0.133	-0.213	**-0.411**	0.469	0.277
X_5	**-0.255**	-0.084	-0.360	0.178	-0.122
X_6	0.202	-0.038	0.189	**-0.181**	0.648
X_7	-0.018	-0.162	0.592	0.022	**-0.372**
X_8	-0.059	-0.269	0.232	0.563	-0.220
X_9	-0.043	-0.049	0.450	0.435	0.408
X_{10}	**0.364**	0.178	-0.012	0.032	-0.216
X_{11}	**0.407**	0.203	0.061	-0.001	-0.058
X_{12}	**0.402**	0.194	0.099	0.021	-0.050
X_{13}	**0.379**	0.156	-0.018	0.104	-0.103
X_{14}	**0.349**	0.050	-0.141	0.296	0.021

Fuente: Dirección nacional de estudios (Base de dato año lectivo 1998–99)

En la tabla CIII, se han resaltado con **negrilla** las variables más importante en cada componente principal con respecto a su matriz de

carga (ver tabla CII), a continuación también se muestran sus combinaciones lineales entre las variables:

I-ésima componente principal

$$Y_1 = -0.215\,X_1 - 0.222\,X_2 - 0.222\,X_3 + \cdots + 0.379\,X_{13} + 0.349\,X_{14}$$

$$Y_2 = +0.554\,X_1 + 0.558\,X_2 + 0.299\,X_3 + \cdots + 0.156\,X_{13} + 0.050\,X_{14}$$

$$Y_3 = -0.003\,X_1 + 0.009\,X_2 + 0.155\,X_3 + \cdots - 0.018\,X_{13} - 0.141\,X_{14}$$

$$Y_4 = +0.233\,X_1 + 0.229\,X_2 - 0.046\,X_3 + \cdots + 0.104\,X_{13} + 0.296\,X_{14}$$

$$Y_5 = -0.037\,X_1 - 0.017\,X_2 + 0.263\,X_3 + \cdots - 0.103\,X_{13} + 0.021\,X_{14}$$