Reconocimiento estadístico de patrones

Guevara Díaz Karlo Jair (Licenciatura en Matemáticas).

Domingo 8 de Mayo del 2022.

1. Introducción

A lo largo del proyecto trabajaremos con datos de municipios de diferentes estados, iniciando con un procesamiento de los datos para posteriormente poder aplicaremos algunos métodos de visualización y de clostering a nuestros datos. Esto con el fin de entender el comportamiento económico de los municipios, encontrar algunas estructuras o patrones entre los municipios, buscar como identificar cuando un municipio esta teniendo problemas económicos (relativamente hablando), descartar algunos métodos de visualización y encontrar gráficas representativas de nuestros datos.

2. Desarrollo

Preprocesamiento de los datos

Daremos una explicación del preprocesa-miento de datos que realizamos y el por que lo preferimos así. Tenemos una gran variedad de características de cada municipio en cada año, por ejemplo los datos para Aguascalientes municipio calientes son lo siguientes (figura 1). 2

POB_TOT	VP	ANALF	SPRIM	OVSDE	OVSEE	OVSAE	VHAC	OVPT	PL<5000	POZSM	OVSD	OVSDSE	IM	GM	INDOA100	LUG_NAC	LUGAR_EST	AÑO
877190	-		2.06	9.54	0.31	0.16	0.72	18.01	0.63	8.73	31.13	-	-		-1.676	Muy bajo		2408	11	2015
797010	-		2.59	11.48	0.49	0.31	0.77	25.25	1.42	8.73	29.9	-	-		-1.768	Muy bajo	7.69	2409	-	2010
723043	-		3.19	13.61	0.77	0.54	1.54	28.86	1.7	8.21	28.37	-	-		-1.831	Muy bajo	-	2419	11	2005
643419	-		3.86	18.04	-	1.12	0.88	32.04	2.22	7.67	37.24	-		1.5	-1.871	Muy bajo	-	2408	11	2000
582827	121790		4.53	-	-	1.62	1.14	-	-	-	-		2.12	-	-1.735	Muy bajo	-	2393	-	1995
506274	-		6.05	27.99	6.55	3.64	3	45.65	5.65	11.47	58.36	-	-		-1.833	Muy bajo	-	2341	9	1990

Figura 1

mas los datos del 2020 que están en otro csv. Algo importante que notamos es que faltan una gran cantidad de datos. Pero también podemos notar que en las categorías CVE_MUN, POB_TOT, ANALF, OVSEE, OVSAE, IM, GM y LUG_NAC no hace falta datos. Por los que nos restringiremos a estos datos, en un subconjunto de los estados, el cual es Aguascalientes, baja california, Baja California Sur, Campeche, Coahuila de Zaragoza y Colima. Notemos que podemos imprimir los datos como series de tiempo ya que cada municipio tiene los datos de 1990,...,2020.

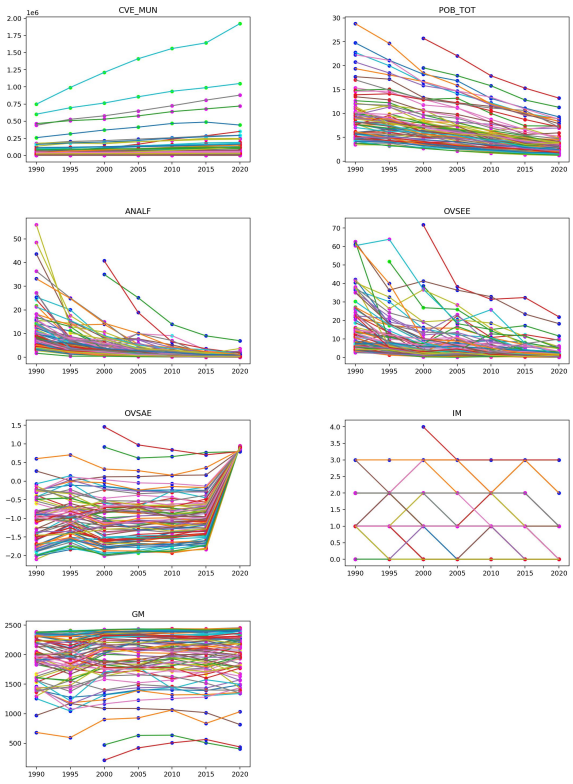

Figura 2

Con esto podemos notar varios municipios no tiene los datos de 1990 y 1995. Por lo que para no tener falta de datos nos restringiremos a los años 2000, 2005, 2010, 2015 y 2020 (podemos hacer un análisis muy similar quitando los municipios donde faltan los primeros años). Ciertamente podríamos iniciar un análisis de los datos como los tenemos ahora (figura 3)

```
[1001.0, 1.0, 877190, 2.06, 0.16, 0.72, -1.676, 0, 2408.0, 2015],
[1001.0, 1.0, 797010, 2.59, 0.31, 0.77, -1.768, 0, 2409.0, 2010],
[1001.0, 1.0, 723043, 3.19, 0.54, 1.54, -1.831, 0, 2419.0, 2005],
[1001.0, 1.0, 643419, 3.86, 1.12, 0.88, -1.871, 0, 2408.0, 2000],
[1001.0, 1.0, 582827, 4.53, 1.62, 1.14, -1.735, 0, 2393.0, 1995],
[1001.0, 1.0, 506274, 6.05, 3.64, 3.0, -1.833, 0, 2341.0, 1990],
```

Figura 3

Donde la primera columna es la clave de cada municipio, la segunda del estado, las demás son los datos antes seleccionados en dicho orden y la ultima el año de los datos. Pero si realizáramos el análisis de datos así estaríamos comparando los municipios en cada año en ves de los municipios en general que es lo que nos interesa en este proyecto. Por lo que cambiaremos la matriz para que en cada fila tenga todos los datos de cada municipio, así en la primera fila tenemos (figura 4).

```
array([948990, 1.64473777347411, 0.113169040426143, 0.378609666467454,
       0.944508383147691, 0, 2435, 948990, 1.64473777347411,
       0.113169040426143, 0.378609666467454, 0.944508383147691, 0, 2435,
       877190, 2.06, 0.16, 0.72, -1.676, 0, 2408.0, 797010, 2.59, 0.31,
       0.77, -1.768, 0, 2409.0, 723043, 3.19, 0.54, 1.54, -1.831, 0,
       2419.0], dtype=object)
```

Figura 4

Así al hacer nuestro análisis de datos estaríamos comparando cada municipio con los demás. En el resto del proyecto usaremos los datos de esta manera.

PCA

Iniciaremos por aplicar PCA a nuestros datos, para tener una primera visualización de los datos. Al realizar esto con dos componentes principales obtenemos la siguiente varianza explicada y componentes principales (figura 5).

```
Pesos  [0.53718741 0.23149371]
PC1  [[-0.14704224  0.16115169  0.18728521  0.20138259 -0.01119777  0.26410975
  -0.06815826 -0.14740276  0.16235057  0.20472105  0.19973523  0.15326553
   0.2254862  -0.06587226 -0.14690706  0.15479695  0.18655889  0.19157904
   0.15430662  0.19279169 -0.06417739 -0.14681421  0.14931955  0.22399846
   0.1692696   0.15628514  0.20628181 -0.0650564  -0.14654754  0.14820386
   0.22482029  0.21643174  0.16852017  0.20884585 -0.06935528]
 [ 0.4066561   0.02174618  0.10286703  0.15635652 -0.00118916  0.11613498
  -0.01350621  0.40379899  0.02123383  0.10816546  0.16670527 -0.00343817
   0.02496476 -0.00707215  0.40490279  0.01977856  0.08777638  0.13880749
  -0.00881384 -0.01877119 -0.00610703  0.40488658  0.01202378  0.11230438
   0.11255978 -0.01795715 -0.0025149  -0.00497081  0.40160068  0.01515412
   0.11654622  0.15832001 -0.00758046  0.00157805 -0.00816221]]
```

Figura 5

Con esto capturamos un 0.768 de la varianza de nuestros datos. Por lo que es una visualización informativa, pero se puede mejorar ya que solo tenemos 0.768 de la varianza. Observemos que el primer componente depende principalmente de las entradas 6, 4, 3 y 2 (GM, OVSAE, OVSEE y ANALF) en ese orden. Por lo que si al hacer la proyección sobre las primeras dos componentes obtenemos un valor grande de la segunda componente significaría que el municipio proyectado tiene un grado de marginación alto (GM), problemas de luz (OVSAE), problemas de agua (OVSEE) y de analfabetismo (ANALF) ya que esos son los calores mas grandes de la primera componente. Por otro lado la segunda componente depende principalmente de las entradas 1 y 3 (POB_TOT y OVSEE) de nuestros datos, lo que significa que un valor grande en la segunda componente significa un gran publicación (POB_TOT) y algo de problemas con el agua (OVSEE). Realizando la proyección sobre las dos componentes y tomando los colores azul para Campeche, verde para Baja California, roza para Coahuila de Zaragoza, amarillo para Aguascalientes, rojo para Colima y azul claro para Baja California Sur obtenemos la siguiente figura (figura 6)

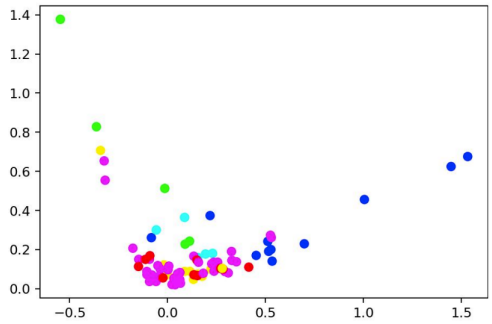

Figura 6

Por lo que tenemos 3 conjuntos distintivos de datos distintivos (figura 7).

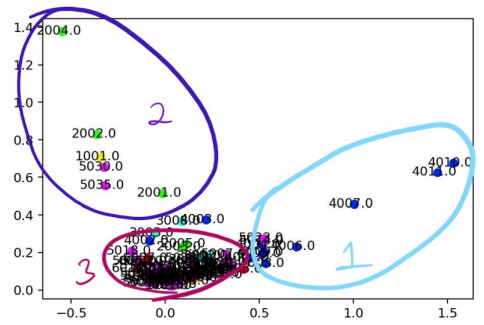

Figura 7

Tenemos que el conjunto 1 son los que tiene un valor en la segunda componente grande y por lo dicho antes serian municipios con claros problemas económicos o de vivienda a comparación de los demás. Por otro lado el conjunto 2 que tiene un primer componente negativo y un segundo componente grande a comparación de los demás tenemos que serian estados con mas población y sin tantos problemas de vivienda y/o pobreza. También esta el conjunto 3 donde estan la mayoria de los municcipios, estos serian los que no tiene un nivel muy alto ni muy bajo de poblacion (a comparaccion de los demas) y su nivel de pobreza no es muy critico como los del conjunto 1.

ISOMAP

Ahora aplicaremos ISOMAP a nuestros datos. Al hacerlo con n = 2 obtenemos (figura 8)

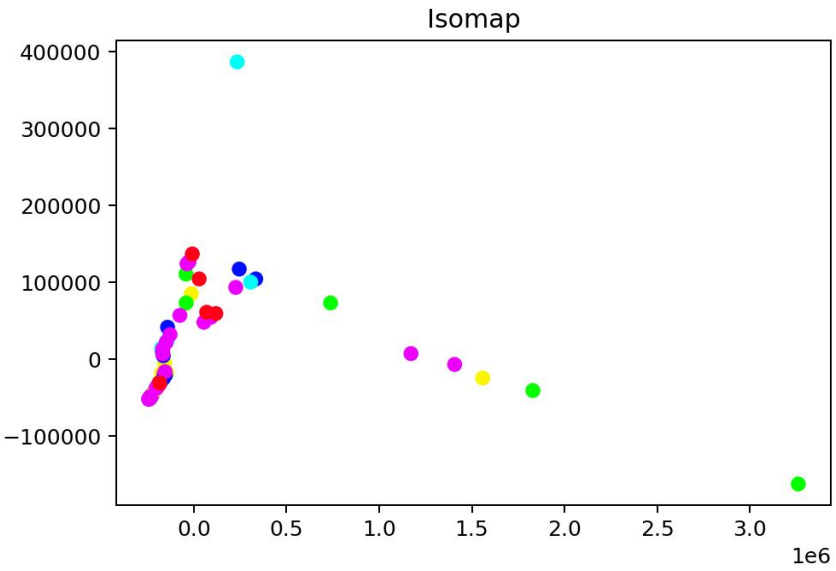

Figura 8

Dado que tenemos $n = 2$, podemos interpretar el resultado como un grafo en una recta (figura 9)

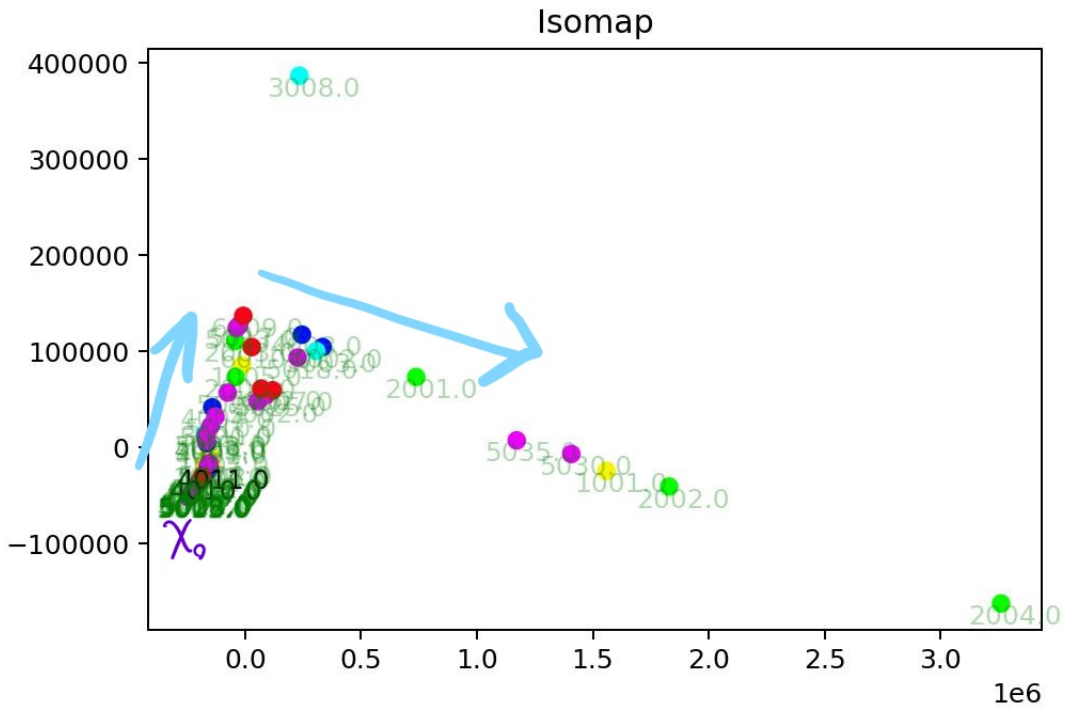

Figura 9

Tenemos en letras negras los municipios 4010 y 4011, los cuales fueron los que en PCA mostraban mas problemas económicos. Podemos notar que el grafo va de x_0 a 2004 o de 4011 a 2004. También podemos notar que a medida que nos alejamos de x_0 tenemos municipios como 2004 (Mexicali), 1001 (Aguascalientes), 3008 (Los cabos) y otros municipios los cuales es conocido que tiene buena economía y dichos municipios son en sus estados los que tiene mejor economías. También es conocido que Campeche (los municipios de campeche tienen etiquetas entre 4000 y 5000 y estan de azul fuerte) tiene problemas económicos ya que su producto interno bruto fue de 1047511.322 en 2013 a 482973.093 en 2020. Por lo que tenemos evidencia a favor para decir que el grafo va de x_0 los municipios con mas problemas económicos a los municipios con menos problemas económicos o mejores economías. También podemos notar que para cada color o estado tenemos puntos alejados del origen x_0, es decir cada estado tiene al menos un municipio con buena economía (incluso campeche aun que por poco). Una posible explicación es que cada estado se enfoca en tener al menos una buena fuente de ingresos para los demás estados (alguien tiene que pagar las cuentas por decirlo de una forma mas coloquial).

T-SNE

Aplicando T-SNE a nuestros datos con 2 componentes, obtenemos (figura 12)

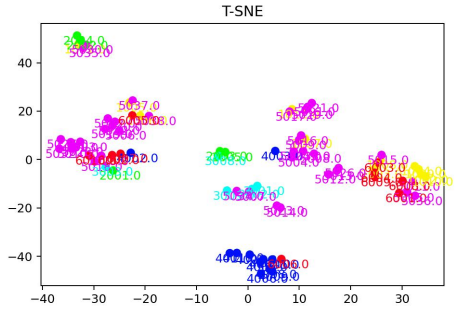

Figura 10

Aun que a primera vista parece que T-SNE no funciono del todo tenemos que en realidad es un muy buen resultado. Por ejemplo veamos los puntos azules (que son los de campeche y también los de etiqueta entre 4000 y 5000), como dijimos antes tenemos que Campeche tiene problemas económicos (su PIB fue de 1047511.322 en 2013 a 482973.093 en 2020) pero también tiene municipios con buena economía en particular 4002 que es campeche y 4003 que es carmen (figura 13)

Municipio	Unidades		Personal ocupado total		Remuneraciones		Producción bruta total		Total de activos fijos	
					Miles de pesos					
	Absoluto	%	Absoluto	%	Absoluto	%	Absoluto	%	Absoluto	%
Total Campeche	**30 022**	**100.0**	**168 919**	**100.0**	**13 691 885**	**100.0**	**720 980 407**	**100.0**	**174 481 603**	**100.0**
Carmen	8 348	27.8	78 445	46.4	11 285 232	82.4	704 160 934	97.7	157 380 066	90.2
Campeche	11 234	37.4	56 291	33.3	1 890 346	13.8	13 364 025	1.9	15 358 973	8.8
Champotón	2 733	9.1	11 008	6.5	159 249	1.2	1 624 365	0.2	527 599	0.3
Calkiní	2 752	9.2	8 026	4.8	107 278	0.8	417 170	0.1	219 990	0.1
Escárcega	1 836	6.1	5 617	3.3	96 827	0.7	664 043	0.1	455 458	0.3
Hecelchakán	890	3.0	2 555	1.5	29 878	0.2	174 872	0.0	83 910	0.0
Candelaria	655	2.2	2 037	1.2	23 982	0.2	162 747	0.0	195 178	0.1
Hopelchén	658	2.2	1 815	1.1	19 521	0.1	119 627	0.0	78 465	0.0
Tenabo	286	1.0	1 278	0.8	55 491	0.4	159 012	0.0	67 410	0.0
Palizada	352	1.2	1 023	0.6	9 268	0.1	42 586	0.0	57 814	0.0
Calakmul	278	0.9	824	0.5	14 813	0.1	91 026	0.0	56 740	0.0

La suma de los porcentajes puede no coincidir con el total debido al redondeo.
Los municipios se ordenaron de acuerdo con la cantidad de personal ocupado total.

Figura 11

Y podemos notar que estos dos municipios están separados de los demás de campeche en el T-SNE lo cual es congrute con lo que queremos ya que el interés no es separar los municipios por estados si no por su economía.
Por otro lado el punto rojo de etiqueta 6006 que esta en la colección de puntos azules corresponde al municipio de Ixtlahuacán del estado de Colima. Este municipio tiene economía y características muy parecidas a las de los municipios de Campeche (como lo son numero de habitantes he ingresos). Un analíticas parecido lo podemos hacer con cada conjunto de puntos, por ejemplo el conjunto de puntos en la esquina superior izquierda $(-30, 45)$ corresponde principalmente a los municipios que hemos dicho que tiene buena economías.
En conclusión el T-SNE separa a los municipios en conjuntos con economías parecidas, esto ya nuestros datos son principalmente sobre características económicas de los municipios, por lo que si se parecen los municipios en estas características al aplicar T-SNE nos dará probabilidades mas grandes y por ende al gráfica esto los puntos quedaran cercanos.

Clustering Jerarquico

Al aplicar " Clustering Jerarquico " a nuestros datos obtenemos el siguiente árbol (figura 10).

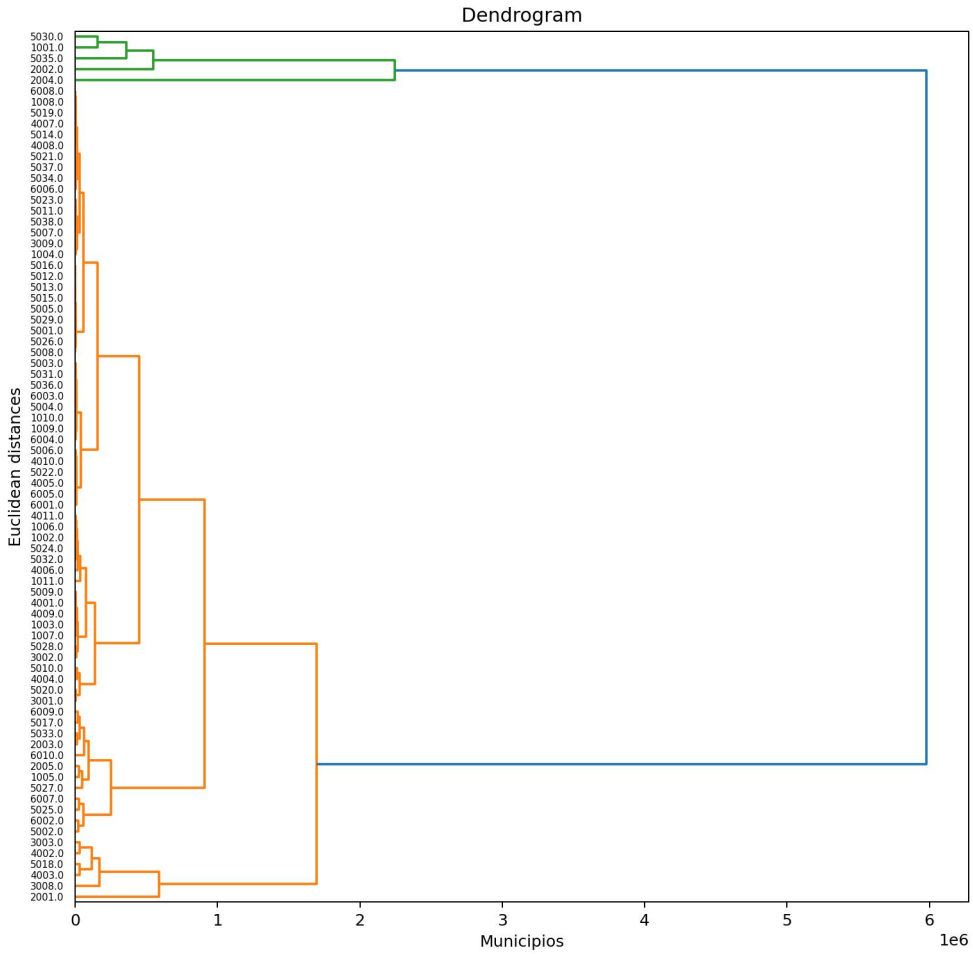

Figura 12

Podemos apreciar varias similitudes con el ISOMAP de antes. Por ejemplo el subarbol verde son los municipios mas alejados del origen x_0 del ISOMAP, por lo que el subarvol verde seria los municipios con buena economía lo que nos indicaría que "Clustering Jerarquico" nos separa muy bien los municipios con buena y mala economía de los demás. También municipios mas cercanos en el árbol también lo están en el ISOMAP. Un ejemplo de lo anterior seria el subárbol conformado por 4011, 1006, 1002, 5024, 5032, 4006 y 1011 que en el ISOMAP esta dado por (figura 11).

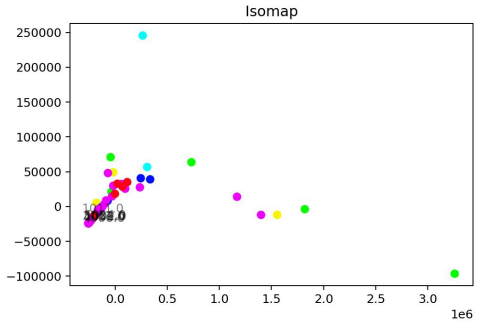

Figura 13

Que son puntos muy cercanos en el ISOMAP. Esto ya que tanto en el ISOMAP y en el árbol usamos la métrica euclidiana. Solo que en el árbol se aprecia mejor como se separan los municipios con economías mas sobresalientes.

SOM

Aplicaremos SOM con los valores $n = 3$ y $m = 2$ para obtener una mejor separación de los grupos y poder compararlo con los otros métodos, obtenemos lo siguiente (figura 14).

```
[4 3 0 0 0 0 0 1 1 1 4 4 3 3 2 0 0 1 5 2 3 0 0 0 1 1 1 1 4 2 2 2 3 3 0 0 3
 0 0 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 3 1 1 1 1 1 1 4 5 1 4
 1 1 1 0 1 1]

array([1001., 1005., 1003., 1011., 1002., 1007., 1006., 1009., 1004.,
       1008., 2002., 2001., 2003., 2005., 3008., 3001., 3002., 3009.,
       4002., 4003., 4004., 4001., 4009., 4011., 4010., 4005., 4007.,
       4008., 5030., 5018., 5025., 5002., 5017., 5033., 5020., 5028.,
       5010., 5009., 5032., 5027., 5006., 5031., 5003., 5004., 5022.,
       5007., 5038., 5011., 5034., 5023., 5014., 5037., 5021., 5008.,
       5029., 5012., 5016., 5005., 5015., 5013., 6007., 6002., 6010.,
       6009., 6001., 6003., 6004., 6008., 6006., 1010., 2004., 3003.,
       4006., 5035., 5019., 5001., 5026., 5024., 5036., 6005.])
```

Figura 14

Revisando el conjunto 0. Tenemos que el conjunto 0 esta conformado por "1003.0, 1011.0, 1002.0, 1007.0, 1006.0, 3001.0, 3002.0, 4001.0, 4009.0, 4011.0, 5020.0, 5028.0, 5009.0, 5032.0,5024.0". Esto coincide con el resultado obtenido en "Clustering Jerarquico" (figura 15). Lo mismo pasa en los de mas conjuntos salvo algunos cambios. Uno de estos cambios es la ubicación del municipio '2001' ya que se coloca en un conjunto diferente al de los municipios con buena economía lo cual no esta mal ya que vimos en otras visualizaciones esta cerca del conjunto de datos con buena economía pero también lo esta del conjunto de datos de mala economía. El inconveniente es que esto no se refleja en el resultado de SOM lo bueno es que si en los en alguno de los demás métodos.

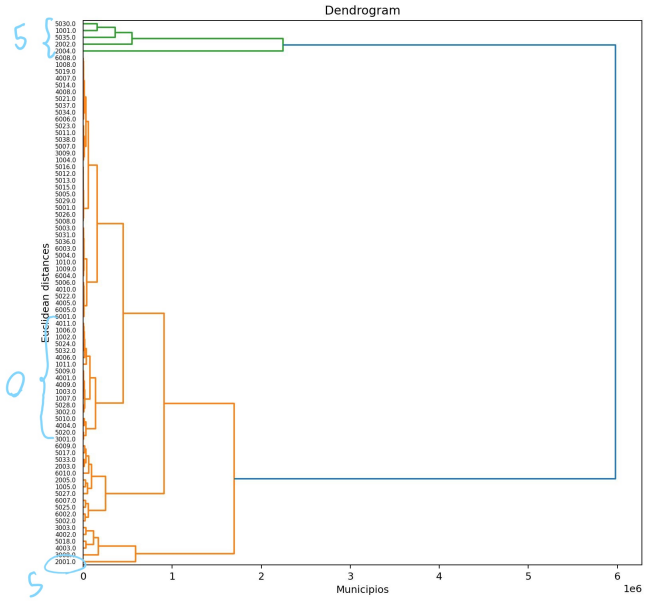

Figura 15

LLE

Pondremos de color azul el texto de los municipios de Campeche (los puntos azul) y de texto verde los municipios que solían estar en el conjunto de municipios con buena economía. Al aplicar LLE con n = 1 a nuestros datos y lo anterior obtenemos (figura 15).

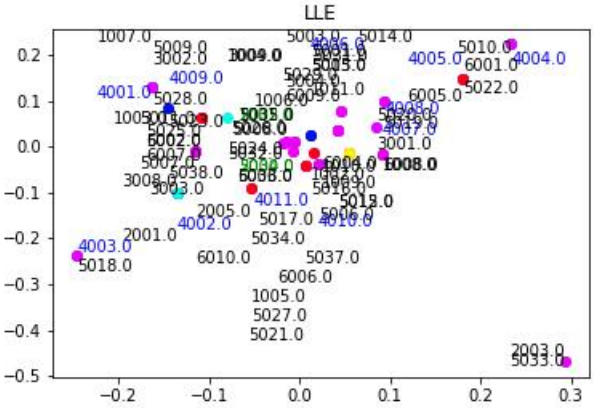

Figura 16

Podemos notar que ninguna de las características que se conservaban con otros métodos se conservan con LLE por ejemplo los municipios en azul que casi siempre se mantenían juntos salvo un par de ellos en este caso se distribuyen en casi todas partes. A lo mucho tenemos que los municipios con buena economía se mantienen cercas pero no se mandan al mismo punto. Al aplicar LLE con $n = 2$, y solo imprimiendo alguno de municipios (para que sea mas claro) obtenemos

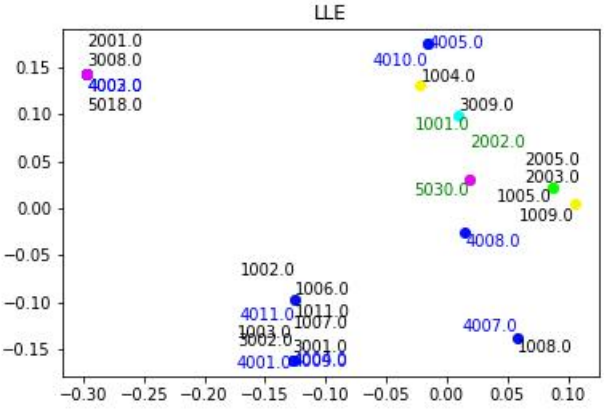

Figura 17

Igual que antes podemos notar que los municipios azules que esperaríamos que se mantengan al menos cercas se distribuyen en casi todas partes y los valores en verde que igual esperaríamos que se mantengan juntos se separan entre 2 puntos. Este resultado poco útil (para nuestros intereses) se puede atribuir al echo de que LLE rescata mas estructuras locales de los datos i.e puede que se mantengan relaciones como $1001 \sim 2002$ o $4007 \sim 1008$ pero no estructuras mas globales como la separación por conjuntos de datos. Esto se podría ajustar un poco tomando un $n = 10$, con lo que obtenemos (figura 18)

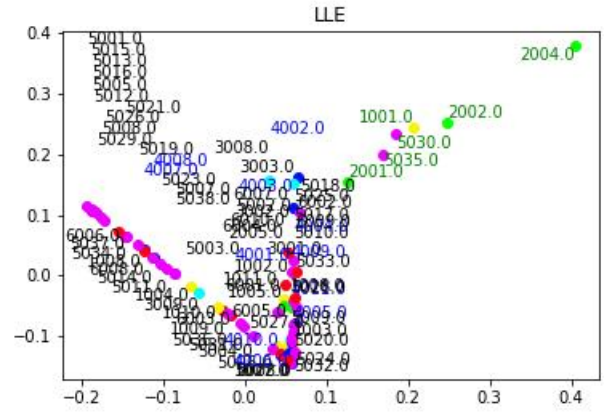

Figura 18

Aun que se obtenemos una ligera mejor, igual tenemos un resultado no tan satisfactorio ya que igual no se distinguen las características que hemos estado manejando o observando. Ciertamente podemos estar rescatando característicos de los datos mas locales pero para nuestros fines no son muy informativas o de alto interés.

Conclusión

Por lo antes explorado descartamos podemos descartar LLE como método de visualización de nuestros datos. Por otro lado al aplicar PCA podemos iniciar a entender el comportamiento de nuestros datos

sobretodo en los valores mas alejados que son los municipios con muy mala o muy buena economía (conjuntos 1 y 2 de la figura 19).

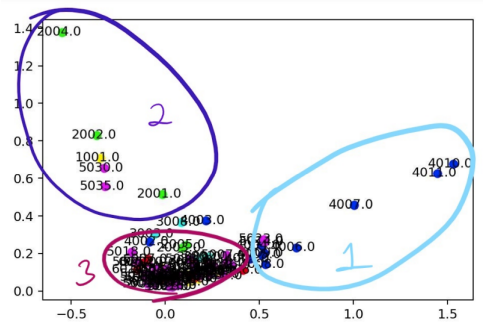

Figura 19

Por otro lado al aplicar ISOMAP a nuestros datos y usando lo aprendido en PCA, pudimos tener una visualización mas global de los datos y apreciar como una "transición"de los municipios con una mala economía a los municipios con una buena economía

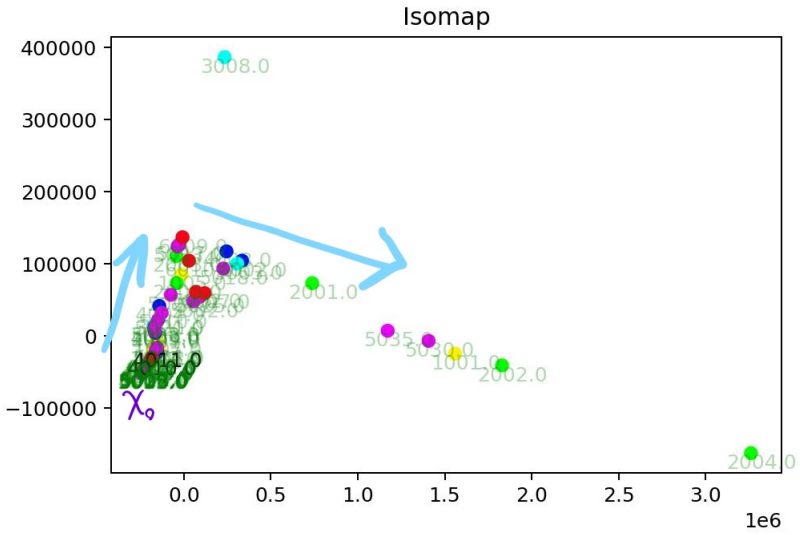

Figura 20

Después al aplicar T-SNE queda mas claro la separación en conjuntos según su economía y se aprecia como se separan algunos municipios (como el de puntos en azul).

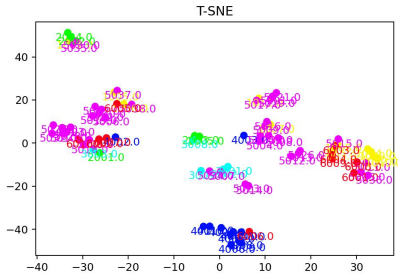

Figura 21

Seguido de los métodos de visualización iniciamos a analizar los métodos de clostering para ver que sea congruente la separaciones que dan con las visualizaciones. Y para estos datos descartamos SOM ya que la información que da esta contenida en Clustering Jerarquico pero Clustering Jerarquico rescata mas información como por ejemplo que el municipio 2001 no esta entre los municipios con mejor economía tampoco esta muy lejos de ellos.

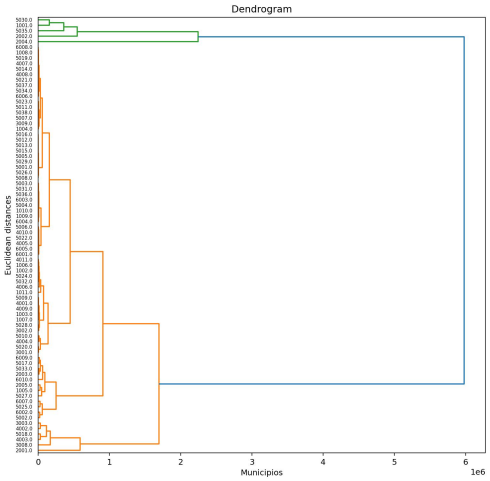

Figura 22

Por lo que aplicando PCA y T-SNE podemos identificar los conjuntos de municipios con problemas económicos y podemos después aplicar Clustering Jerarquico para comprara resultados y evitar dar aproximaciones falsas.