

**DETECCIÓN DE ZONAS DE POTENCIAL CRECIMIENTO PARA
LAS MICRO EMPRESAS EN NUEVO LEÓN, NAYARIT Y
YUCATÁN, BASADO EN TÉCNICAS DE *Clustering*
ESPECTRAL**

T E S I S

Que para obtener el grado de
Maestra en Cómputo Estadístico

Presenta

Yareli Aleidali Macías Ángeles

Director de Tesis:

Dra. Graciela Ma. González Farías

Autorización de la versión final

A mi familia, por todo su apoyo y por ser una inspiración en mi camino.

Resumen

En México, las micro empresas representan alrededor de 95 % de los establecimientos y concentran a 37.2 % de la población ocupada, por lo que representan el primer canal por el cual muchos mexicanos cuentan con un trabajo o deciden emprender y ser fuente de trabajo.

En este proyecto de tesis se presenta una propuesta de modelo de *clustering* que permita detectar las zonas que han presentado un escenario favorecedor para los micro negocios, es decir, en las cuales se ha dado un crecimiento económico acompañado de bienestar social y no sólo como consecuencia de la expansión urbana.

El modelo toma como base información pública del INEGI a nivel localidad, del DENUE se obtiene la información sobre los negocios en México y del Censo se obtienen las variables socio-demográficas. Posteriormente, se aplica la técnica de selección de variables SPEC propuesto por [Zhao y Liu. \(2007\)](#), de la cual se obtiene un *ranking* por nivel de relevancia de cada variable y que es utilizado en la fase agrupamiento.

Se aplica el algoritmo de *Clustering Espectral* propuesto por [Ng, Jordan, y Weiss \(2002\)](#) con un par de variantes: (1) en la función de similitud RBF kernel se utiliza un parámetro de escala local σ_i en lugar del parámetro global σ [Zelnik-Manor y Perona \(2004\)](#) y (2) se elige el valor óptimo de clústeres utilizando criterios de ajuste clásicos en conjunto con el heurístico Eigengap y la Búsqueda iterativa del Eigengap [Afzalan y Jazizadeh \(2019\)](#).

El resultado de este trabajo se presenta en forma de mapas interactivos de consulta pública, los cuales contienen un *ranking* de localidades por estado de acuerdo al potencial del sector micro y al nivel de bienestar que han presentado en los últimos años.

Palabras clave: aprendizaje máquina no supervisado, *Clustering Espectral*, grafos, DENUE, Censo, micro empresas, crecimiento, mapas.

Agradecimientos

Al Centro de Investigación en Matemáticas Unidad Monterrey, por permitirme regresar a la vida académica y continuar con mi formación, agradezco a cada uno de los profesores por su gran labor a lo largo de la maestría y en especial, con gran cariño, a la Dra. Graciela por acompañarme durante la elaboración de este trabajo así como por todos sus consejos.

Al Consejo Nacional de Ciencia y Tecnología por haberme proporcionado apoyo económico durante el programa de maestría.

Índice general

Resumen	III
Agradecimientos	V
Índice de figuras	IX
Índice de tablas	XIII
1. Introducción	1
1.1. Antecedentes	1
1.2. Definición del problema y Objetivo	2
2. Fuentes de datos, preproceso y selección de variables	7
2.1. Fuentes de datos	7
2.1.1. DENUE	7
2.1.2. Censo de Población y Vivienda	12
2.1.3. Marco Geoestadístico	13
2.2. Pre-procesamiento de la información y obtención de variables	15
2.2.1. DENUE	15
2.2.2. Censo de Población y Vivienda	17
2.3. Método de selección de variables para <i>clustering</i>	19
2.3.1. Spectral Feature Selection (SPEC)	21
3. Modelo	31
3.1. <i>Clustering</i> Espectral	31
3.1.1. Grafo de Similitud	34
3.1.2. Matriz Laplaciana de un grafo	36
3.1.3. Algoritmos de <i>Clustering</i> Espectral	38
3.1.4. Función de similitud RBF kernel y parámetro de escala local σ^2	40
3.1.5. Número de clústeres	44
4. Modelación	51
4.1. Nayarit	51
4.1.1. Análisis exploratorio	51
4.1.2. Selección de variables mediante algoritmo SPEC.	55
4.1.3. <i>Clustering</i> Espectral	63

VIII *ÍNDICE GENERAL*

4.2.	Nuevo León	94
4.2.1.	Análisis exploratorio	94
4.2.2.	Selección de variables mediante algoritmo SPEC.	96
4.2.3.	<i>Clustering</i> Espectral.	97
4.3.	Yucatán	117
4.3.1.	Análisis exploratorio	117
4.3.2.	Selección de variables mediante algoritmo SPEC.	118
4.3.3.	<i>Clustering</i> Espectral	119
5.	Conclusiones y trabajo futuro	141
	Referencias	145
A.	Implementaciones y mapa interactivo	149

Índice de figuras

1.1.	Tipos de establecimientos en México por tamaño. Censo Económico 2019 - Resultados oportunos. (INEGI, 2019c)	2
1.2.	Tipo de establecimientos en México por tamaño, personal ocupado y valor agregado. Resultados definitivos. (INEGI, 2019c)	3
1.3.	Proporción de nacimientos y muertes de establecimientos a nivel nacional, por tamaño de establecimiento. (INEGI, 2020b)	3
1.4.	Vocación de las entidades según su Valor Agregado Censal Bruto. (INEGI, 2019c)	5
2.1.	Diagrama RENEM INEGI. (2018)	8
2.2.	Estructura establecimientos Censo Económico 2019	10
2.3.	Sectores a considerar en el modelo para el estado de Nayarit.	10
2.4.	Sectores a considerar en el modelo para el estado de Nuevo León.	11
2.5.	Sectores a considerar en el modelo para el estado de Yucatán.	11
2.6.	Relevancia de las características Aggarwal y Reddy (2014).	20
2.7.	Caso donde el corte mínimo provee una mala partición.	25
2.8.	Grafos con 1, 2 y 4 componentes independientes Fleshman (2019).	28
3.1.	Conjuntos de datos con formas complejas Liu y Han (2014).	32
3.2.	Regiones cóncavas y convexas.	32
3.3.	Ejemplo dona. <i>K-means</i> vs. <i>Clustering</i> Espectral	33
3.4.	Kernel polinomial de grado 2 sobre datos en \mathbb{R}^2	41
3.5.	Función RBF bajo distintos parámetros de σ^2 Sreenivasa (2020).	42
3.6.	<i>Clustering</i> Espectral sin parámetro de escala local Zelnik-Manor y Perona (2004).	42
3.7.	Efecto del parámetro de escala local Zelnik-Manor y Perona (2004).	43
3.8.	Tres conjuntos de datos y los 10 eigenvalores más pequeños de \mathcal{L}_{rw} Luxburg (2007).	47
3.9.	Diagrama Búsqueda Iterativa del Eigengap.	49
4.1.	Información general Nayarit	52
4.2.	Localidades con Unidades Económicas	52
4.3.	UE 2020. Localidades por ámbito.	53
4.4.	Distribución localidades Nayarit. Cluster -1, 0 y Otros.	54
4.5.	Dispersión variables originales. Base DENUE 126 localidades Nayarit.	56

4.6. Resultados algoritmo SPEC con parámetro local σ_i . Base DENUE 126 localidades Nayarit estandarizada.	57
4.7. Localidades sin variables socio-demográficas.	58
4.8. Dispersión variables originales. Base CPV 122 localidades Nayarit. . .	59
4.9. Resultados algoritmo SPEC con parámetro local σ_i . Base CPV 122 localidades Nayarit estandarizada.	60
4.10. Resultados algoritmo SPEC con parámetro local σ_i . Base DENUE-CPV 122 localidades Nayarit estandarizada.	62
4.11. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base de 12 variables DENUE, 126 localidades Nayarit estandarizada.	64
4.12. Resultados <i>clustering</i> espectral con parámetro local σ_i . Base de 12 variables y 126 localidades Nayarit estandarizada.	65
4.13. Distribución de clústeres. Base de 12 variables y 126 localidades Nayarit estandarizada.	66
4.14. Resultados al eliminar las variables de menor relevancia Base DENUE 126 localidades Nayarit estandarizada.	67
4.15. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base de 5 variables DENUE, 126 localidades Nayarit estandarizada.	68
4.16. Resultados <i>clustering</i> espectral con parámetro local σ_i . Base de 5 variables y 126 localidades Nayarit estandarizada.	69
4.17. Distribución de clústeres. Base de 5 variables y 122 localidades Nayarit estandarizada..	70
4.18. Eigensearch.	71
4.19. Dispersión clústeres DENUE y cuantiles .25, .50 y .75 poblacionales. .	72
4.20. Matriz de correlación variables originales vs. <i>embedding</i>	73
4.21. Dim1 vs. Dim2	74
4.22. Dim3 vs. Dim4	74
4.23. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base de 22 variables CPV, 122 localidades Nayarit estandarizada.	79
4.24. Resultados <i>clustering</i> espectral con parámetro local σ_i . Base de 22 variables del CPV y 122 localidades Nayarit estandarizada.	80
4.25. Distribución de clústeres por variable según <i>ranking</i> de relevancia. Base de 22 variables del CPV y 122 localidades Nayarit estandarizada. . .	81
4.26. Resultados <i>clustering</i> eliminando las variables de menor relevancia Base CPV 122 localidades Nayarit estandarizada.	82
4.27. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base de 7 variables CPV, 122 localidades Nayarit estandarizada.	83
4.28. Resultados <i>clustering</i> espectral con parámetro local σ_i . Base de 7 variables del CPV y 122 localidades Nayarit estandarizada.	84
4.29. Distribución de clústeres por variable según <i>ranking</i> de relevancia. Base de 22 variables del CPV y 122 localidades Nayarit estandarizada. . .	85
4.30. Eigensearch.	86
4.31. Dispersión clústeres CPV y cuantiles .25, .50 y .75 poblacionales. .	87
4.32. Matriz de correlación variables originales vs. <i>embedding</i>	88
4.33. Dim1 vs. Dim2	89

4.34. Dim3 vs. Dim4	89
4.35. Localidades Nayarit por categoría.	92
4.36. Localidades con mayor potencial de crecimiento y bienestar social Nayarit.	92
4.37. <i>Ranking</i> por municipio Nayarit.	93
4.38. Información general Nuevo León	94
4.39. Localidades con Unidades Económicas	94
4.40. Distribución localidades Nuevo León. Cluster -1, 0 y Otros.	95
4.41. Resultados algoritmo SPEC con parámetro local σ_i . Base DENUE y base CPV Nuevo León estandarizada.	96
4.42. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base de 12 variables DENUE, 113 localidades Nuevo León estandarizada.	97
4.43. <i>clustering</i> $k = 4$, $k = 3$ y $k = 2$. Base de 12 variables DENUE, 113 localidades Nuevo León estandarizada.	98
4.44. Resultados al eliminar las variables de menor relevancia. Base DENUE 113 localidades Nuevo León estandarizada.	98
4.45. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base de 9 variables DENUE, 113 localidades Nuevo León estandarizada.	99
4.46. Resultados <i>clustering</i> espectral $k = 4$ con parámetro local σ_i . Base de 9 variables y 113 localidades Nuevo León estandarizada.	99
4.47. Eigensearch.	100
4.48. Dispersión clústeres DENUE y cuantiles .25, .50 y .75 poblacionales. .	101
4.49. Matriz de correlación variables originales vs. <i>embedding</i>	102
4.50. Dim1 vs. Dim2	102
4.51. Dim3 vs. Dim4	103
4.52. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base de 22 variables CPV, 98 localidades Nuevo León estandarizada.	106
4.53. Resultados al eliminar las variables de menor relevancia. Base CPV 98 localidades Nuevo León estandarizada.	107
4.54. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base de 13 variables CPV, 98 localidades Nuevo León estandarizada.	107
4.55. Resultados <i>clustering</i> espectral $k = 3$. Base 13 variables y 98 localidades Nuevo León estandarizada.	108
4.56. Eigensearch.	108
4.57. Dispersión clústeres CPV y cuantiles .25, .50 y .75 poblacionales. .	109
4.58. Matriz de correlación variables originales vs. <i>embedding</i>	110
4.59. Dim1 vs. Dim2	111
4.60. Dim2 vs. Dim3	111
4.61. Localidades Nuevo León por categoría.	114
4.62. Localidades con mayor potencial de crecimiento micro y bienestar social Nuevo León.	114
4.63. <i>Ranking</i> por municipio Nuevo León.	116
4.64. Información general Yucatán	117
4.65. Localidades con Unidades Económicas Yucatán	117
4.66. Distribución localidades Yucatán. Cluster -1, 0 y Otros.	118

4.67. Resultados algoritmo SPEC con parámetro local σ_i . Base DENUE y base CPV Yucatán estandarizada. <i>Ranking</i> de mayor a menor relevancia.	119
4.68. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base de 12 variables DENUE, 127 localidades Yucatán estandarizada.	120
4.69. <i>clustering</i> $k = 4$, $k = 3$ y $k = 2$. Base de 12 variables DENUE, 127 localidades Yucatán estandarizada.	121
4.70. Resultados al eliminar las variables de menor relevancia. Base DENUE 127 localidades Yucatán estandarizada.	121
4.71. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base 5 variables DENUE, 127 localidades Yucatán estandarizada.	122
4.72. Resultados <i>clustering</i> espectral $k = 4$ con parámetro local σ_i . Base de 5 variables y 127 localidades Yucatán estandarizada.	122
4.73. Eigensearch.	123
4.74. Dispersión clústeres DENUE y cuantiles .25, .50 y .75 poblacionales. .	124
4.75. Matriz de correlación variables originales vs. <i>embedding</i>	124
4.76. Dim1 vs. Dim2	125
4.77. Dim3 vs. Dim4	126
4.78. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base de 22 variables CPV, 127 localidades Yucatán estandarizada.	129
4.79. Resultados al eliminar las variables de menor relevancia. Base CPV 127 localidades Yucatán estandarizada.	130
4.80. Grafo de similitudes, top 5 eigengaps y <i>embedding</i> espectral. Base de 13 variables CPV, 127 localidades Yucatán estandarizada.	131
4.81. Resultados <i>clustering</i> espectral $k = 3$. Base 13 variables y 127 localidades Yucatán estandarizada.	131
4.82. Eigensearch.	132
4.83. Dispersión clústeres CPV y cuantiles .25, .50 y .75 poblacionales. .	133
4.84. Matriz de correlación variables originales vs. <i>embedding</i>	134
4.85. Dim1 vs. Dim2	135
4.86. Dim2 vs. Dim3	135
4.87. Localidades Yucatán por categoría.	137
4.88. Localidades con mayor potencial de crecimiento micro y nivel intermedio de bienestar social Yucatán.	138
4.89. Localidades con mayor nivel de bienestar social y nivel medio de crecimiento del sector micro.	139
4.90. Ranking por municipio Yucatán.	140
A.1. Ejemplo Dashboard Nayarit	149

Índice de tablas

2.1. Dimensiones bases DENUE Nayarit 2010-2020	8
2.2. Comparativo bases de datos del Censo de Población y Vivienda 2020	13
2.3. Descripción bases de datos del Marco Geoestadístico INEGI (2021). .	14
4.1. Distribución localidades Nayarit. <i>Cluster -1, Cluster 0 y Otros.</i>	53
4.2. Clústeres DENUE Nayarit.	72
4.3. <i>Ranking</i> Clústeres DENUE Nayarit.	77
4.4. Clústeres CPV Nayarit.	86
4.5. <i>Ranking</i> Clústeres CPV.	91
4.6. Distribución localidades Nuevo León. <i>Cluster -1, 0 y Otros.</i>	95
4.7. Clústeres DENUE Nuevo León.	100
4.8. <i>Ranking</i> Clústeres DENUE Nuevo León.	105
4.9. Clústeres CPV Nuevo León.	109
4.10. <i>Ranking</i> Clústeres CPV Nuevo León.	113
4.11. Distribución localidades Yucatán. <i>Cluster -1, 0 y Otros.</i>	118
4.12. Clústeres DENUE Yucatán.	123
4.13. <i>Ranking</i> Clústeres DENUE Yucatán.	128
4.14. Clústeres CPV Yucatán.	132
4.15. <i>Ranking</i> Clústeres DENUE Yucatán.	137

Capítulo 1

Introducción

1.1. Antecedentes

Durante mucho tiempo ha existido el interés por medir el crecimiento económico de las economías, se han creado índices económicos y se han desarrollado diversas teorías tratado de explicar las variables que impulsan el desarrollo en las economías. En este sentido, se puede afirmar que se trata de un tema bastante estudiado desde diversas perspectivas y que a través de los años se ha venido haciendo un esfuerzo por recopilar datos que midan desde distintos enfoques el crecimiento de las economías, prueba de ello son los censos económicos que realiza el Instituto Nacional de Estadística y Geografía (INEGI) cada 5 años a partir de 1989.

Ante un tema que contempla tantas variables para su medición, es posible que se tienda a estudiar de forma descriptiva y por separado cada una de sus variables dejando de lado la visión integral que requiere el problema. Por tal motivo, se observa la necesidad de crear modelos estadísticos que logren reflejar la información de las distintas variables a través de elementos claros y concisos que sirvan a los usuarios para la toma de decisiones. Bajo esta lógica, en este trabajo de tesis se presenta una propuesta para identificar las zonas de potencial crecimiento para un sector en particular, las micro empresas.

1.2. Definición del problema y Objetivo

En México, las micro empresas, definidas como aquellos establecimientos que cuentan con un tamaño de personal de 0 a 10, juegan un papel de gran importancia en la economía ya que, de acuerdo con los últimos tres censos económicos (2009, 2014, 2019), alrededor del 95 % de los establecimientos (aprox. 6 millones en 2019) son de esta clase.

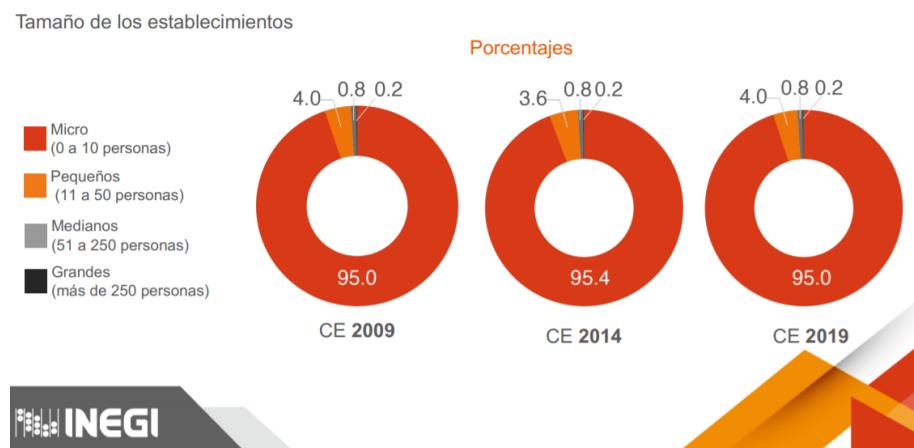


Figura 1.1: Tipos de establecimientos en México por tamaño. Censo Económico 2019 - Resultados oportunos. (INEGI, 2019c)

Así mismo, en 2019 el 37.2 % de la población ocupada trabajaba para las micro empresas, porcentaje superior a las pequeñas y medianas empresas PyME's (de 11 a 250 empleados) con 30.7 % y grandes empresas (más de 250 empleados) con 32.1 %. Además, cabe resaltar que el 24 % de los micro negocios correspondieron a emprendimientos ya que llevaban menos de dos años de vida INEGI (2019c).

A pesar de la gran cantidad de micro negocios en nuestro país y de su importante contribución al empleo, se tiene evidencia de que éstos concentran el menor porcentaje de ingresos, medido a partir del Valor Agregado Censal Bruto (sólo 14.6 %), en comparación con las pequeñas, medianas y grandes empresas INEGI (2019b).

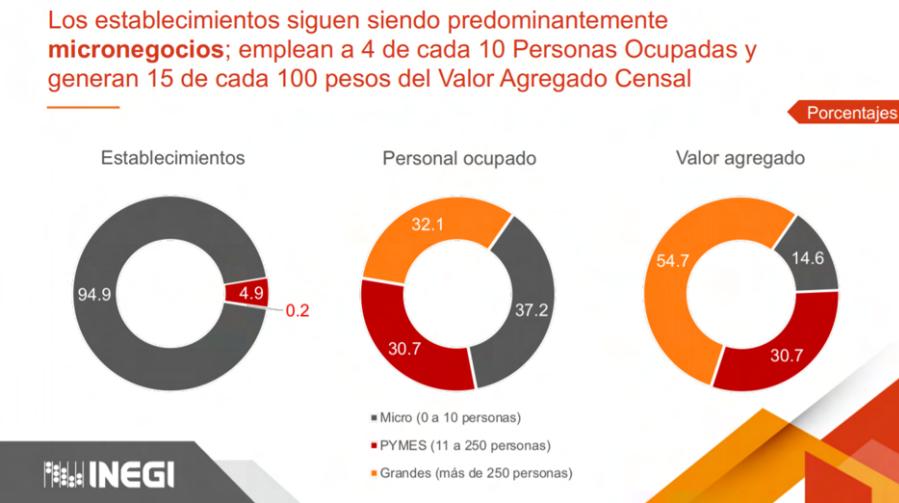


Figura 1.2: Tipo de establecimientos en México por tamaño, personal ocupado y valor agregado. Resultados definitivos. (INEGI, 2019c)

Aunado a lo anterior, el Estudio sobre la Demografía de los Negocios 2020 del INEGI, revela que a 17 meses de haberse concluido el censo económico 2019 (en Septiembre 2020) y en el contexto de la pandemia por Covid-19, aproximadamente 20.8 % de los micro negocios registrados en el censo de 2019 murieron pero a la vez hubo un 13 % de nacimientos de los mismos, ésta última cifra cobra importancia si se compara con las PyME's donde sólo el 2.8 % corresponde a nacimientos de negocios.

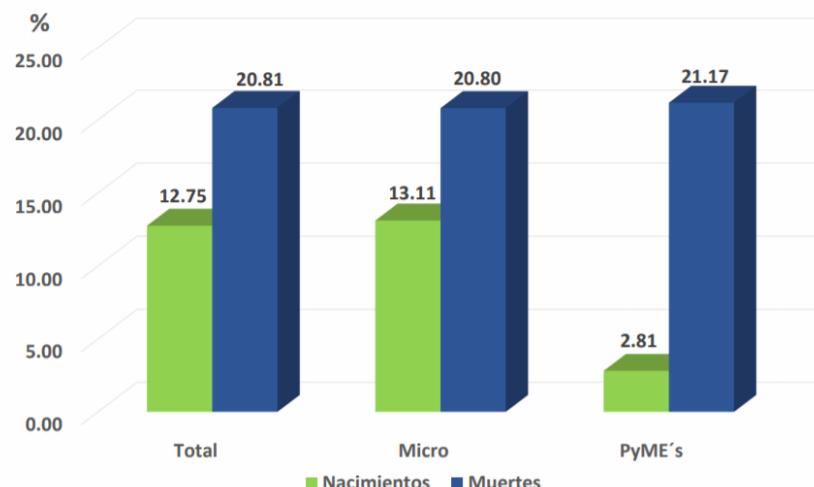


Figura 1.3: Proporción de nacimientos y muertes de establecimientos a nivel nacional, por tamaño de establecimiento. (INEGI., 2020b)

Por lo anterior, en este trabajo de tesis se presenta una propuesta de modelo de *clustering* que permita detectar las zonas que de acuerdo con los últimos años han presentado un escenario favorecedor para las micro empresas en México, en las cuales se ha registrado un aumento en el número de negocios de este tipo y que en algunos casos ha habido transformaciones de micro a pequeña, mediana e incluso grandes empresas. Es de interés identificar aquel crecimiento económico que viene acompañado de bienestar social y no sólo como consecuencia de la expansión urbana. Es por ello, que se propone incluir en el modelo variables socio-demográficas tales como la tasa de desempleo, el grado de escolaridad promedio, nivel de analfabetismo, entre otras.

El resultado de este trabajo se presenta en forma de mapas interactivos de consulta pública, los cuales contienen un *ranking* de localidades por estado de acuerdo al potencial del sector micro y al nivel de bienestar que han presentado en los últimos años. La intención es que esta herramienta sirva de guía para todo aquél que este interesado en emprender o invertir en un micro negocio en México, ya que a través de éstos es posible generar empleos, lo cual se traduce en una oportunidad de mejorar las condiciones de vida de las personas y en consecuencia la del país propio.

El estudio se realizará para tres estados de la República Mexicana: Nuevo León, Nayarit y Yucatán. La razón por la cuál se eligieron estos estados (uno correspondiente al norte, otro al centro y otro al sur del país) es debido a que son economías distintas entre si, donde se tienen diversas formas de vida y condiciones sociales. Como se puede observar en la Figura 1.4, en Nuevo León el sector con mayor presencia es el correspondiente a actividades industriales, mientras que Nayarit tiene mayor presencia en el sector Servicios y Yucatán en el sector Comercio.

El modelo toma como base información pública del Instituto Nacional de Estadística y Geografía (INEGI) a nivel localidad; del Directorio Estadístico Nacional de Unidades Económicas(DENUE) se obtiene la información sobre los negocios en México y del Censo de Población y Vivienda (CPV) se obtienen las variables socio-demográficas. En el Capítulo 2 se describen aspectos como la temporalidad de los datos, los sectores económicos a considerar y la descripción de las variables. Pos-



Figura 1.4: Vocación de las entidades según su Valor Agregado Censal Bruto. (INEGI, 2019c)

teriormente, se desarrolla la parte teórica de la técnica de Selección de Variables Espectral conocida como SPEC, con el este algoritmo se obtiene un *ranking* de las variables por nivel de relevancia, el cual se utilizará en el proceso de agrupamiento.

El agrupamiento de las observaciones se obtendrá aplicando un modelo de *Clustering* Espectral. En el Capítulo 3 se describen los diferentes algoritmos de este tipo de *clustering*, sus elementos y los parámetros que lo componen, así como algunas variantes que se han propuesto con la finalidad de mejorar la calidad de los clústeres.

En el Capítulo 4 se expone el proceso y los resultados de realizar el agrupamiento para los tres estados, se inicia con un análisis exploratorio, se muestran los resultados de la técnica de selección de variables y se describe el proceso que se siguió para determinar los grupos finales, bajo un escenario donde se consideran todas las variables y un escenario donde se eliminan las variables de menor relevancia. Por último, se realiza la interpretación de los grupos generados de acuerdo a las dispersiones de los clústeres y con base en esto se asigna una puntuación a cada clúster, de manera que el resultado final sea un *ranking* a nivel localidad en el cual un mayor puntaje indicará un mejor escenario para las micro empresas.

Los resultados se presentan en forma de mapas interactivos de consulta pública, en los cuales se presenta el *ranking* de las localidades por estado de acuerdo al potencial

del sector micro y al nivel de bienestar que han presentado en los últimos años, por último se incluye el detalle de los micro negocios geo-referenciados por sectores económicos.

Capítulo 2

Fuentes de datos, preproceso y selección de variables

2.1. Fuentes de datos

2.1.1. DENU

El Directorio Estadístico Nacional de Unidades Económicas (DENU) es una herramienta que publica el INEGI desde julio 2010, la cual recopila información sobre los negocios activos en México. Con ella es posible identificar a las Unidades Económicas (UE) por su nombre comercial, por tipo de organización jurídica (personas físicas o morales), por su actividad económica, por su tamaño (con base en la cantidad de personal ocupado) y además brinda las coordenadas geográficas para poder ubicar los establecimientos en el territorio nacional [INEGI. \(2020a\)](#).

Para llevar a cabo el presente trabajo, se recopilaron todas las actualizaciones del DENU desde 2010 hasta 2020 teniendo un total de 16 ediciones. Es importante mencionar que las primeras ediciones del DENU (2010, 2011, 2012, 2013_07 y 2013_10) presentan algunas variaciones entre ellas ya que difieren en nombre y número de campos, por lo que hasta 2015 se tienen bases homogéneas.

Cabe señalar que en 2015, se crea el Registro Estadístico de Negocios de México (RENEM), instrumento integrador de las diferentes fuentes de información internas y

8 CAPÍTULO 2. FUENTES DE DATOS, PREPROCESO Y SELECCIÓN DE VARIABLES

Periodo	Dimensión
2010	(46282,24)
2011	(46827,30)
2012	(47088,47)
2013-07	(47324,44)
2013-10	(47324,44)
2015	(53812,41)
2016-01	(57090,41)
2016-10	(57387,41)
2017-03	(57415,41)
2017-11	(57501,41)
2018-03	(57508,41)
2018-11	(57577,41)
2019-04	(57905,41)
2019-11	(65895,41)
2020-04	(66270,41)
2020-11	(67027,41)

Tabla 2.1: Dimensiones bases DENUE Nayarit 2010-2020

externas de las UE y del cual el DENUE es su vista pública. La creación del RENEM brindó mayor formalidad al DENUE, por lo cual en este trabajo de tesis se consideró adecuado utilizar la información del DENUE a partir de 2015, ya que las bases se encuentran homogéneas y mejor consolidadas¹.

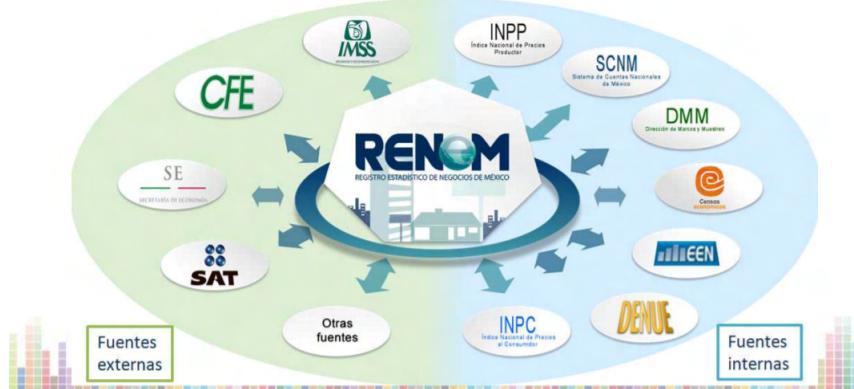


Figura 2.1: Diagrama RENEM INEGI. (2018)

La actualización del DENUE se realiza de manera **total** cada 5 años a través de los

¹Cabe mencionar que la base de DENUE de 2010 se utilizará sólo para saber cuántas UE se encontraban dadas de alta en ese año y poder comparar con la información disponible de 2015 en adelante.

censos económicos (2009, 2014, 2019) y de forma **parcial** cada año en dos formas 1) los negocios grandes y ciertos sectores económicos se actualizan mediante información proveniente de registros administrativos de las Unidades del Estado (IMSS, CFE, SAT, etc.) y de las Encuestas Económicas Nacionales realizadas por el INEGI, 2) el segmento de negocios micro, pequeños y medianos se actualiza de manera parcial a través de registros administrativos; y en ambos subuniversos los usuarios pueden realizar actualizaciones de manera continua a través de la herramienta de actualización del DENUE Interactivo [INEGI. \(2018\)](#).

Por último, es importante señalar que el DENUE contempla los establecimientos de todos los sectores económicos clasificados a código de 6 dígitos en el Sistema de Clasificación Industrial de América del Norte 2018 (SCIAN 2018), sin embargo, para fines de este trabajo de tesis se decide sólo considerar los establecimientos pertenecientes a los sectores con mayor presencia y fortaleza económica para el estado de Nayarit, Nuevo León y Yucatán.

Para poder identificar estos sectores se cuantificó a nivel Sector² el número de establecimientos y el Valor Agregado Censal Bruto (VACB)³. Debido a que el DENUE no contiene información acerca del VACB fue necesario obtener estas medidas a través de los datos abiertos del Censo Económico de 2019 [INEGI \(2019a\)](#).

Las cifras del VACB obtenidas del Censo Económico 2019 corresponden a un subconjunto del universo de establecimientos, es decir, sólo contempla establecimientos que iniciaron operaciones antes de 2019, ubicados en la zona urbana y que forman parte del Sector Privado y Paraestatal (Ver Figura 2.2), no obstante, como la mayoría de establecimientos se encuentran dentro de este subconjunto, las cifras obtenidas son un buen indicador.

²Sector es la clasificación a dos dígitos de acuerdo al catálogo SCIAN 2018

³Valor agregado censal bruto (millones de pesos): Es el valor de la producción que se añade durante el proceso de trabajo por la actividad creadora y de transformación del personal ocupado, el capital y la organización (factores de la producción), ejercida sobre los materiales que se consumen en la realización de la actividad económica. Aritméticamente, el Valor Agregado Censal Bruto Total (VACB) resulta de restar a la Producción Bruta Total el Consumo Intermedio. Se le llama bruto porque no se le ha deducido el consumo de capital fijo [INEGI \(2019a\)](#)

10 CAPÍTULO 2. FUENTES DE DATOS, PREPROCESO Y SELECCIÓN DE VARIABLES



Figura 2.2: Estructura establecimientos Censo Económico 2019

Para el estado de Nayarit se contemplaron 10 sectores. Se eligieron aquellos que tenían mayor presencia, medida por el porcentaje de unidades económicas (UE) respecto al total en el estado y también se consideran los sectores con mayor VACB por UE, el cual resulta de dividir el VACB entre el número de UE. De ésta manera también se considerarán aquellos sectores que no tienen tanta presencia pero que su VACB es significativo, como es el caso del sector Construcción (Ver Figura 2.5).

Código SCIAN	Sector	UE	%	VACB	%	VACB por UE	Dentro del modelo
11	Agricultura, cría y explotación de animales, aprovechamiento forestal, pesca y caza	1,322	2%	\$642	2%	0.49	
21	Minería	29	0%		0%	0.00	
22	Generación, transmisión, distribución y comercialización de energía eléctrica, suministro de agua y de gas natural por ductos al consumidor final	38	0%		0%	0.00	
23	Construcción	294	1%	\$820	2%	2.79	*
31-33	Industrias manufactureras	5,436	10%	\$3,442	10%	0.63	*
43	Comercio al por mayor	1,358	2%	\$4,034	11%	2.97	*
46	Comercio al por menor	21,283	37%	\$11,444	32%	0.54	*
48-49	Transportes, correos y almacenamiento	211	0%	\$619	2%	2.93	*
51	Información en medios masivos	86	0%	\$176	0%	2.05	*
52	Servicios financieros y de seguros	341	1%	\$1,545	4%	4.53	*
53	Servicios inmobiliarios y de alquiler de bienes muebles e intangibles	1,008	2%	\$733	2%	0.73	
54	Servicios profesionales, científicos y técnicos	1,066	2%	\$602	2%	0.56	
55	Corporativos	2	0%		0%	0.00	
56	Servicios de apoyo a los negocios y manejo de residuos, y servicios de remediación	680	1%	\$2,781	8%	4.09	*
61	Servicios educativos	530	1%	\$582	2%	1.10	
62	Servicios de salud y de asistencia social	2,381	4%	\$602	2%	0.25	
71	Servicios de esparcimiento culturales y deportivos, y otros servicios recreativos	647	1%	\$210	1%	0.32	
72	Servicios de alojamiento temporal y de preparación de alimentos y bebidas	11,545	20%	\$6,703	19%	0.58	*
81	Otros servicios excepto actividades gubernamentales	8,766	15%	\$1,197	3%	0.14	*
Total general		57,023	100%	\$36,132	100%	0.63	

Figura 2.3: Sectores a considerar en el modelo para el estado de Nayarit.

Para el estado de Nuevo León se contemplaron 13 sectores (Ver Figura 2.4).

Código SCIAN	Sector	UE	%	VACB	%	VACB por UE	Dentro del modelo
11	Agricultura, cría y explotación de animales, aprovechamiento forestal, pesca y caza	40	0%	\$66	0%	1.64	
21	Minería	42	0%	\$775	0%	18.46	*
22	Generación, transmisión, distribución y comercialización de energía eléctrica, suministro de agua y de gas natural por ductos al consumidor final	15	0%	\$36,891	4%	2459.42	*
23	Construcción	1,188	1%	\$18,092	2%	15.23	*
31-33	Industrias manufactureras	14,001	9%	\$358,752	41%	25.62	*
43	Comercio al por mayor	7,620	5%	\$67,919	8%	8.91	*
46	Comercio al por menor	57,193	38%	\$91,622	10%	1.60	*
48-49	Transportes, correos y almacenamiento	1,356	1%	\$32,071	4%	23.65	*
51	Información en medios masivos	351	0%	\$9,300	1%	26.50	*
52	Servicios financieros y de seguros	1,575	1%	\$97,201	11%	61.71	*
53	Servicios inmobiliarios y de alquiler de bienes muebles e intangibles	3,104	2%	\$11,171	1%	3.60	
54	Servicios profesionales, científicos y técnicos	4,668	3%	\$16,470	2%	3.53	
55	Corporativos	72	0%	\$30,035	3%	417.15	*
56	Servicios de apoyo a los negocios y manejo de residuos, y servicios de remediación	2,862	2%	\$46,626	5%	16.29	*
61	Servicios educativos	2,421	2%	\$18,006	2%	7.44	
62	Servicios de salud y de asistencia social	7,466	5%	\$8,461	1%	1.13	
71	Servicios de esparcimiento culturales y deportivos, y otros servicios recreativos	1,681	1%	\$5,788	1%	3.44	
72	Servicios de alojamiento temporal y de preparación de alimentos y bebidas	19,997	13%	\$13,543	2%	0.68	*
81	Otros servicios excepto actividades gubernamentales	25,796	17%	\$9,993	1%	0.39	*
Total general		151,448	100%	\$872,782	100%	5.76	

Figura 2.4: Sectores a considerar en el modelo para el estado de Nuevo León.

Para el estado de Yucatán se contemplaron 9 sectores (Ver Figura 2.4).

Código SCIAN	Sector	UE	%	VACB	%	VACB por UE	Dentro del modelo
11	Agricultura, cría y explotación de animales, aprovechamiento forestal, pesca y caza	1,264	1%	\$1,081	1%	0.85	
21	Minería	49	0%	\$279	0%	5.69	
22	Generación, transmisión, distribución y comercialización de energía eléctrica, suministro de agua y de gas natural por ductos al consumidor final	112	0%	\$4,304	4%	38.42	*
23	Construcción	609	1%	\$4,338	4%	7.12	*
31-33	Industrias manufactureras	26,715	24%	\$21,319	21%	0.80	*
43	Comercio al por mayor	3,127	3%	\$21,153	21%	6.76	*
46	Comercio al por menor	38,225	34%	\$22,557	22%	0.59	*
48-49	Transportes, correos y almacenamiento	377	0%	\$3,328	3%	8.83	*
51	Información en medios masivos	210	0%	\$1,180	1%	5.62	
52	Servicios financieros y de seguros	849	1%	\$1,471	1%	1.73	
53	Servicios inmobiliarios y de alquiler de bienes muebles e intangibles	1,714	2%	\$1,823	2%	1.06	
54	Servicios profesionales, científicos y técnicos	2,145	2%	\$2,626	3%	1.22	
55	Corporativos	5	0%	\$81	0%	16.18	*
56	Servicios de apoyo a los negocios y manejo de residuos, y servicios de remediación	1,432	1%	\$5,022	5%	3.51	
61	Servicios educativos	1,399	1%	\$2,416	2%	1.73	
62	Servicios de salud y de asistencia social	4,153	4%	\$1,488	1%	0.36	
71	Servicios de esparcimiento culturales y deportivos, y otros servicios recreativos	1,455	1%	\$525	1%	0.36	
72	Servicios de alojamiento temporal y de preparación de alimentos y bebidas	15,331	14%	\$5,282	5%	0.34	*
81	Otros servicios excepto actividades gubernamentales	13,332	12%	\$2,176	2%	0.16	*
Total general		112,503	100%	\$102,448	100%	0.91	

Figura 2.5: Sectores a considerar en el modelo para el estado de Yucatán.

2.1.2. Censo de Población y Vivienda

El INEGI ofrece de manera pública dos bases de datos que incluyen 222 indicadores sobre las características de la población y de las viviendas, *Principales resultados por localidad (ITER)* y *Principales resultados por AGEb y manzana urbana*.

La base *ITER* presenta los indicadores del Censo de Población y Vivienda (CPV) de todas las localidades en el país, aunque cabe señalar que para las localidades que contienen una y dos viviendas se presenta la información de manera agrupada y sólo proporciona la información de los indicadores Población total (POBTOT), Viviendas totales (VIVTOT) y Total de viviendas habitadas (TVIVHAB), el resto de los indicadores aparecen asteriscos (nulos) [INEGI \(2020b\)](#).

La segunda base, *Principales resultados por AGEb y manzana urbana*, proporciona los 222 indicadores del CPV para las localidades urbanas⁴ del país, desagregados hasta el nivel de área geoestadística básica⁵ (AGEB) y manzana urbana [INEGI \(2020a\)](#).

Para fines de este trabajo de tesis se decide utilizar la base *ITER*. Como se puede apreciar en la Tabla 2.2, poniendo como ejemplo al Estado de Nayarit, si se utilizan los datos a nivel localidad se logra abarcar 2,850 localidades (urbanas y rurales, amanzanadas y puntuales) de las 4,902 existentes en el marco geoestadístico del estado , esto representa el 60 por ciento del total ya que el resto son localidades con 1 o dos viviendas de las cuales no se tiene la información desagregada.

Por el contrario, si se toma la base a nivel AGEb y manzana, se puede ser más específico en cuanto a la división de las zonas geográficas, sin embargo, se deja fuera una gran cantidad de localidades debido a que no tienen más de 2500 habitantes o no son cabeceras municipales, y por ejemplo, para este estado se perderían localidades como Nuevo Vallarta la cual alberga a una gran cantidad de establecimientos.

⁴Una localidad urbana es aquella que tiene una población mayor o igual a 2,500 habitantes o que es rural, es decir, tiene menos de 2500 habitantes pero que es cabecera municipal [INEGI \(2020a\)](#).

⁵Un área geoestadística básica (AGEB) es la extensión territorial que corresponde a la subdivisión de las áreas geoestadísticas municipales. Dependiendo de sus características, se clasifican en dos tipos: AGEb urbana y AGEb rural. Un área geoestadística básica urbana, es un área geográfica ocupada por un conjunto de manzanas perfectamente delimitadas por calles, avenidas, andadores o cualquier otro rasgo de fácil identificación en el terreno y cuyo uso del suelo es principalmente habitacional, industrial, de servicios, comercial, etcétera, y sólo son asignados al interior de las localidades urbanas [INEGI \(2020a\)](#).

Concepto	Total Nayarit	CPV por Loc	CPV por ageb_mza
POBTOT	1,235,456	1,235,456	895,565
MUN-LOC	4,902	2,850	58

Tabla 2.2: Comparativo bases de datos del Censo de Población y Vivienda 2020

Con base en lo anterior, se procedió a descargar de manera automática los indicadores del CPV 2010 y 2020 a nivel localidad para el estado de Nayarit, Nuevo León y Yucatán.

Es importante mencionar que se identificaron localidades que en 2010 existían bajo cierto nombre y que en la actualidad ya no existen, esto generará una ligera perdida de información del 2010, sin embargo, es un porcentaje mínimo (0.7 % aprox.) que para fines de este trabajo no representa un impacto en los resultados.

2.1.3. Marco Geoestadístico

El Marco Geoestadístico muestra la división geoestadística del territorio nacional en sucesivos niveles de desagregación y tiene cobertura en todas las localidades de México. Este se compone de varios archivos los cuales corresponden a las capas de desagregación de las entidades. A continuación se especifican los niveles que lo componen.

Base	Descripción
eeent	Áreas geoestadísticas estatales
eemun	Áreas geoestadísticas municipales
ear	Áreas geoestadísticas básicas rurales
eel	Polígono de localidades urbanas y rurales amanzanadas
eelp	Localidades puntuales rurales
eeti	Territorio insular
ea	Áreas geoestadísticas básicas urbanas
eem	Polígonos de manzana
eefm	Frentes de manzana
eee	Ejes de vialidad
ecd	Caserío disperso
esia	Servicios e información complementaria de Tipo área
esil	Servicios e información complementaria de Tipo línea
esip	Servicios e información complementaria de Tipo puntual
epe	Polígono externo
eepam	Polígono externo de manzana

Tabla 2.3: Descripción bases de datos del Marco Geoestadístico INEGI (2021).

En este caso utilizaremos las capas **eel** y **eelp**⁶ con las cuales ubicaremos los polígonos de las localidades urbanas y rurales, y además, las localidades rurales puntuales (localidades rurales que no están amanzanadas). Con la unión de ambos subconjuntos cubriremos todas las localidades del territorio.

⁶“ee” corresponde a la clave numérica de la entidad federativa: 01, 02,... 32. Cabe aclarar que las capas con sufijo ti, cd, pe, pem, sia, sil, sip, se incluyen únicamente si la localidad cuenta con este tipo de información.

2.2. Pre-procesamiento de la información y obtención de variables

Para llevar a cabo la unión de la información obtenida a través de las distintas fuentes y los distintos periodos, se consideró conveniente tomar como base las localidades de más de 2 viviendas reportadas en el CPV 2020 y sobre ésta unir el resto de información, esto se realizó debido a que se identificaron localidades que en 2010 existían bajo cierto nombre y que en la actualidad ha cambiado, por lo que habrá perdida de información, sin embargo, es un porcentaje mínimo (0.7 % aprox.)

Cabe mencionar que la unión de la información se realizó en lenguaje Python, el código generado puede ser consultado a través del enlace que se anexa en el Apéndice [A](#).

2.2.1. DENUE

Como se mencionó anteriormente, se concatenó la información de 2015 a 2020 del DENUE, y aparte se consideró la base DENUE 2010, a partir de esas dos bases se obtuvieron los siguientes indicadores:

1. **UE_2010_perhab.** UE micro en 2010 por habitante de acuerdo al CPV.
2. **UE_2020_perhab.** UE micro en 2020 por habitante de acuerdo al CPV.
3. **Sob UE 2010.** Unidades micro dadas de alta en 2010 que sobreviven en 2020 (de acuerdo al último DENUE) dividido entre la unidades micro existentes en 2010.
4. **UE_Muere_x_cada2020.** Unidades micro que murieron entre el periodo 2015 - abril 2020 dividido entre el número de unidades vivas en noviembre 2020.
5. **UE_IncC_viva2020.** Unidades micro vivas en noviembre 2020 que registraron en algún momento un aumento de personal y conservaron ese aumento dividido entre el total de unidades micro vivas en noviembre 2020

6. **UE_viva2020_Recien.** Unidades micro vivas en noviembre 2020 que se dieron de alta a partir de 2019 (de recién creación) dividido entre el total de unidades micro vivas en noviembre 2020.
7. **UE_viva2020_Media.** Unidades micro vivas en noviembre 2020 que se dieron de alta entre 2014 y 2018 (de media creación) dividido entre el total de unidades micro vivas en noviembre 2020.
8. **UE_viva2020_Larga.** Unidades micro vivas en noviembre 2020 que se dieron de alta entre 2010 a 2013 (de larga creación) dividido entre el total de unidades micro vivas en noviembre 2020.
9. **UE_IncC_peq_viva2020.** Unidades micro vivas en noviembre 2020 que registraron en algún momento un aumento de personal, pasando de ser micro a pequeña empresa, y conservaron ese aumento dividido entre el total de unidades micro vivas en noviembre 2020.
10. **UE_IncC_med_viva2020.** Unidades Micro vivas en noviembre 2020 que registraron en algún momento un aumento de personal, pasando de ser micro a mediana empresa, y conservaron ese aumento dividido entre el total de unidades micro vivas en noviembre 2020.
11. **UE_IncC_gde_viva2020.** Unidades micro vivas en noviembre 2020 que registraron en algún momento un aumento de personal, pasando de ser micro a empresa grande, y conservaron ese aumento dividido entre el total de unidades micro vivas en noviembre 2020.
12. **UE_DecC_viva2020.** Unidades micro vivas en noviembre 2020 que registraron en algún momento un decremento de personal, de tener de 6 a 10 personas a 0 a 5 personas contratadas, y conservaron ese decremento dividido entre el total de unidades micro vivas en noviembre 2020.

2.2.2. Censo de Población y Vivienda

Se descargaron las bases *ITER* de los últimos dos censos 2010 y 2020. Cabe mencionar que en este último se tienen 222 indicadores mientras que en el de 2010 se tienen 190 indicadores, tomando esto en consideración, se eligieron las siguientes variables para incluir en el modelo, a continuación se describen [INEGI \(2020b\)](#):

1. **PROM_HNV.** Promedio de hijas e hijos nacidos vivos 2010/2020. Resultado de dividir el total de hijas e hijos nacidos vivos de las mujeres de 12 a 130 años de edad, entre el total de mujeres del mismo grupo de edad. Excluye a las mujeres que no especificaron el número de hijas e hijos nacidos vivos y a las que sí han tenido, pero no especificaron el total de ellos.
2. **GRAPROES.** Grado promedio de escolaridad 2010/2020. Resultado de dividir el monto de grados escolares aprobados por las personas de 15 a 130 años de edad entre las personas del mismo grupo de edad. Excluye a las personas que no especificaron los grados aprobados.
3. **t_PNACOE.** Población nacida en otra entidad 2010/2020 dividido entre el total de personas en la localidad.
4. **t_P8A14AN** Población de 8 a 14 años que no sabe leer y escribir dividido entre el total de personas en ese grupo de edad, 2010/2020.
5. **t_P15YM_AN.** Población de 15 años y más analfabeta dividido entre el total de personas en ese grupo de edad, 2010/2020.
6. **t_P15YM_SE.** Población de 15 años y más sin escolaridad, que no aprobaron ningún grado escolar o que sólo tienen nivel preescolar dividido entre el total de personas en ese grupo de edad, 2010/2020.
7. **t_PDESOCUP.** Población de 12 años y más desocupada, no tenían trabajo, pero buscaron trabajo en la semana de referencia, dividido entre la población de 12 años y más económicamente activa, 2010/2020.

8. **t_PSINDER.** Población sin afiliación a servicios de salud, en ninguna institución pública o privada, dividido entre el total de personas en la localidad, 2010/2020.
9. **t_VPH_PISOTI.** Total de viviendas particulares habitadas con piso de tierra⁷ dividido entre el total de viviendas particulares habitadas de cualquier clase incluyendo a las viviendas particulares sin información de ocupantes, 2010/2020.
10. **t_VPH_C_SERV.** Viviendas particulares habitadas que disponen de energía eléctrica, agua entubada de la red pública y drenaje dividido entre el total de viviendas particulares habitadas de cualquier clase incluyendo a las viviendas particulares sin información de ocupantes, 2010/2020.
11. **VPH_SNBIEN.** Viviendas particulares habitadas sin ningún bien (no cuentan con refrigerador, lavadora, horno de microondas, automóvil o camioneta, motocicleta o motoneta, bicicleta que se utilice como medio de transporte, algún aparato o dispositivo para oír radio, televisor, computadora, laptop o tablet, Internet, línea telefónica fija, teléfono celular, servicio de televisión de paga (cable o satelital), servicio de películas, música o videos de paga por Internet, ni consola de videojuegos), dividido entre el total de viviendas particulares habitadas de cualquier clase incluyendo a las viviendas particulares sin información de ocupantes, 2010/2020.

⁷VPH_PISOTI, VPH_C_SERV y VPH_SNBIEN, comprenden las viviendas particulares para las que se captaron las características de la vivienda, clasificadas como: casa única en el terreno; casa que comparte terreno con otra(s); casa dúplex; departamento en edificio; vivienda en vecindad o cuartería; vivienda en cuarto de azotea de un edificio y no especificado de vivienda particular. Incluye a las viviendas particulares sin información de ocupantes INEGI (2020b)

2.3. Método de selección de variables para *clustering*

De acuerdo con Salem Alelyani, Jiliang Tang y Huan Liu en [Aggarwal y Reddy \(2014\)](#), ante una creciente cantidad de datos que se producen en la actualidad, el etiquetado manual realizado por los humanos se ha vuelto una tarea extremadamente cara y difícil de realizar, de ahí la necesidad e importancia de realizar etiquetados automáticos como parte indispensable de la minería de datos, la técnica de etiquetado más popular es el *clustering*.

Dado un conjunto de observaciones, la técnica de *clustering* busca formar grupos o clústeres de observaciones de tal forma que dentro del clúster las observaciones sean lo más similares posibles y que entre los distintos clústeres sean lo más disímiles posible. Ésta técnica se utiliza en varias tareas de aprendizaje automático y minería de datos, incluida la segmentación de imágenes, recuperación de información (information retrieval), reconocimiento de patrones (pattern recognition), análisis de grafos, etc. Hoy en día existen múltiples técnicas de *clustering* en la literatura los cuales pueden clasificarse de manera general en método de partición, métodos probabilísticos, métodos jerárquicos, métodos basados en densidad, métodos espectrales, entre otros.

Una de las preguntas que surgen al implementar técnicas de *clustering* es ¿Qué variables ayudan a discriminar mejor las observaciones?, dicha cuestión se responderá a lo largo de esta sección.

De acuerdo con [Aggarwal y Reddy \(2014\)](#), la fase de selección de características (o variables) es un paso que forma parte del pre-procesamiento el cual es importante realizar para mejorar la calidad del aprendizaje, disminuir el costo computacional y facilitar la interpretación del modelo. No todas las variables serán igual de relevantes para encontrar los grupos (aprendizaje no supervisado) o para predecir la variable *target* (aprendizaje supervisado), ya que algunos pueden ser más ruidosos que otros. Varios estudios han demostrado que las variables irrelevantes pueden ser removidas sin que esto deteriore el desempeño del *clustering* [Zhao y Liu. \(2007\)](#).

En la Figura 2.6 se puede apreciar el caso en donde una variable es o no relevante

al momento de formar los clústeres, en este caso la variable f_1 presenta un mayor poder discriminativo con respecto a f_2 y f_3 .

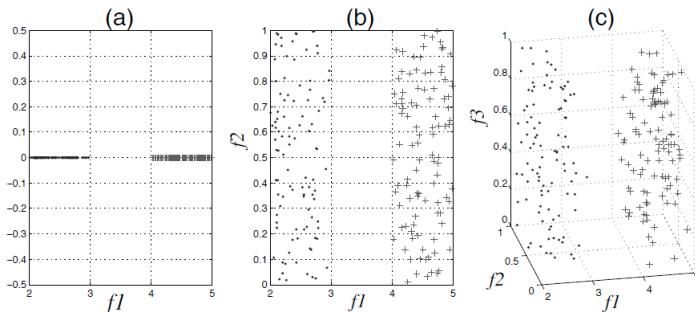


Figura 2.6: Relevancia de las características Aggarwal y Reddy (2014).

Debido a que la finalidad de este trabajo es proponer un modelo de *clustering* que forme grupos de localidades similares de acuerdo a ciertas características económicas y demográficas, se trata entonces de un problema perteneciente al aprendizaje no supervisado, es decir, no se conocen a priori los grupos o etiquetas de los datos.

En el aprendizaje supervisado es relativamente fácil definir cuáles son las características que son relevantes dado que se conoce la variable *target*. Por el contrario, los datos sin etiquetar plantean otro desafío en la selección de variables pues en esos casos es difícil definir la relevancia, sin embargo, aplicar un método de selección de variables puede ayudar a mejorar el aprendizaje de forma similar a como se hace en el aprendizaje supervisado.

Los algoritmos de selección de características se dividen en 3 grupos de acuerdo a si la información se encuentra etiquetada: supervisados cuando se tiene la etiqueta a priori, no supervisados cuando no se conoce la etiqueta previamente o semisupervisados cuando se conoce la etiqueta de una pequeña porción de los datos. A su vez, estos se dividen de acuerdo a la estrategia de selección en: *filter model*, *wrapper model*, *embedded model* e *hybrid model*. Los modelos tipo filtro evalúan la relevancia de las características utilizando algún criterio estadístico independientemente del clasificador que se utilice (aprendizaje supervisado) o de la técnica de *clustering* (aprendizaje no supervisado). Los modelos *wrapper* utilizan el clasificador o técnica de *clustering* como criterio de selección, ya que seleccionan las características que tienen el mayor

poder discriminante medido por el ajuste del clasificador o la calidad del *clustering*. Los modelos híbridos, son una combinación de los dos anteriores ya que primero aplican un criterio estadístico de selección (modelo tipo filtro) y posteriormente se elige el subconjunto de variables que lograron obtener el mayor ajuste al utilizar el clasificador o técnica de *clustering*. Finalmente, los modelos de *embedding* realizan la selección de características mientras se realiza el aprendizaje, es decir, ajusta el modelo y selecciona las características simultáneamente (este último sólo aplica para el aprendizaje supervisado).

Los algoritmos de tipo filtro son preferibles por su bajo costo computacional y debido a que no existe sesgo al elegir un método de *clustering* previamente, no obstante, si se conoce previamente la técnica de *clustering* a aplicar entonces los modelos *wrapper* presentan mejores resultados. Los modelos híbridos son el punto intermedio en cuanto a eficiencia y calidad de *clustering* en comparación con los modelos filtro y *wrapper* Aggarwal y Reddy (2014).

En el presente trabajo se aborda el algoritmo de selección de características para aprendizaje no supervisado de tipo filtro Spectral Feature Selection conocido como SPEC.

2.3.1. Spectral Feature Selection (SPEC)

El algoritmo SPEC funciona tanto para aprendizaje supervisado como no supervisado para problemas de clasificación y esta basado en la teoría de grafos espectral.

Zhao y Liu. (2007) han demostrado que algoritmos de gran eficiencia como ReliefF (aprendizaje supervisado) y Laplacian Score (aprendizaje no supervisado) son casos particulares del algoritmo SPEC, por lo que en 2007 presentan el algoritmo SPEC, el cual es la primer propuesta de algoritmo en unificar la selección de características para el caso supervisado como no supervisado para poder estudiarlo en un marco general.

La similitud entre pares de instancias u observaciones se ha utilizado ampliamente en el aprendizaje supervisado y no supervisado para representar las relaciones entre instancias. La separabilidad de las observaciones puede ser estudiada mediante del

análisis del espectro del grafo generado por el conjunto de similitudes entre pares de instancias S . Por tanto, al desarrollar un algoritmo que determina la relevancia de las características con base en la matriz de similitudes S será posible utilizarlo tanto para aprendizaje supervisado como no supervisado.

Para describir el algoritmo se define la siguiente nomenclatura:

- Sea X el conjunto de n observaciones con m dimensiones, $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ con $\mathbf{x}_i \in R^m$.
- F_1, F_2, \dots, F_m son las m variables o características y $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m$ sus correspondientes vectores de características.
- $Y = (y_1, y_2, \dots, y_n)$ son las etiquetas de clase, sólo aplica para el caso supervisado.
- S es la matriz que guarda las similitudes entre pares de instancias. Para el caso no supervisado [Zhao y Liu. \(2007\)](#) proponen utilizar la función RBF (Radial Basis Function) kernel o kernel Gaussiano.

$$s_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}} \quad (2.1)$$

, donde el parámetro σ controla el ancho de los vecindad de las observaciones. Cabe resaltar que en el siguiente capítulo se presenta una propuesta para elegir el parámetro adecuado de manera local para cada par de observaciones.

Para el caso supervisado la matriz de similitud estará dada por:

$$s_{ij} = \begin{cases} \frac{1}{n_l}, & y_i = y_j = l \\ 0, & \text{otherwise} \end{cases} \quad (2.2)$$

donde n_l es el número de instancias en la clase l .

- Dado el conjunto de observaciones X , $\mathbb{G}(V, E)$ denota el grafo no dirigido construido a partir de S , donde V es el conjunto de nodos que en este caso son las instancias $\mathbf{x}_i \in X$ y E son las aristas o puentes que conectan los pares de nodos (v_i, v_j) , cuyo peso w_{ij} estará determinado por S , es decir $w_{ij} = s_{ij}$. Por lo

tanto, dado el grafo $\mathbb{G}(V, E)$, la matriz de adyacencia o afinidad W esta definida como $W(i, j) = w_{ij}$ para $i \neq j$ y $W(i, i) = 0$ para indicar que el nodo no esta conectado consigo mismo.

- D es la matriz de grado del grafo \mathbb{G} definida como $D(i, j) = d_i$ si $i = j$ y 0 en otro caso, donde $d_i = \sum_{j=1}^n w_{ij}$. Es decir, D es una matriz diagonal y cada d_i es la suma de similitudes de la i -ésima instancia con respecto al resto de observaciones por lo que d_i puede ser interpretado como una estimación de la densidad alrededor de \mathbf{x}_i , mientras haya más observaciones similares o cerca de \mathbf{x}_i , el grado de la instancia i será mayor.
- L es la matriz Laplaciana del grafo \mathbb{G} y \mathcal{L} la matriz Laplaciana normalizada de \mathbb{G} definidas de la siguiente manera:

$$L = D - W \quad (2.3)$$

$$\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \quad (2.4)$$

La matriz Laplaciana satisface algunas propiedades importantes que se especificarán con detalle en el Capítulo 3.

Una variable que es consistente con la estructura del grafo, el cual fue construido a partir de las similitudes, asignará valores similares a las instancias que están cercanas entre si en el grafo. De acuerdo a la teoría de grafos, la información de la estructura de un grafo puede ser obtenida a partir de su espectro, y con base en este medir la relevancia de las variables

De manera general, la idea detrás del algoritmo es estimar la relevancia de las variables por medio de la estimación de la consistencia de la variable con el espectro de la matriz de similitudes S .

La relevancia de cada vector de características \mathbf{f}_i es evaluada usando tres funciones: φ_1 , φ_2 y φ_3 . Estas funciones se derivan de la función de Corte Normalizado (Normalized Cut) propuesto por [Shi y Malik \(2000\)](#).

Corte Normalizado

[Shi y Malik \(2000\)](#) presentan una propuesta con enfoque hacia el campo de la segmentación de imágenes, en la cual se identifican los distintos segmentos en la imagen por medio de la identificación de grupos dentro de un grafo no dirigido, cuyos nodos son los puntos en el espacio de características (pixeles) y las aristas entre cada par de nodos corresponden a una función de similitud.

El objetivo consistía en particionar los nodos en subconjuntos cuya similitud es alta dentro del mismo grupo y baja entre distintos grupos. La pregunta a la que se enfrentaron era ¿Qué criterio utilizar para particionar el grafo de manera eficiente?. [Shi y Malik \(2000\)](#) desarrollaron un nuevo criterio basado en teoría de grafos que mide la calidad de una partición de imágenes, a lo cual llamaron Normalized Cut o Corte Normalizado.

En teoría de grafos se denomina corte (cut) a la suma de los pesos de las aristas que son removidas al particionar un grafo en dos grupos. Por ejemplo, un grafo $\mathbb{G}(V, E)$ puede ser particionado en dos conjuntos disjuntos A y B , donde $A \cup B = V$ y $A \cap B = \emptyset$, al remover las aristas que conectan ambos conjuntos, de manera que:

$$cut(A, B) = \sum_{u \in A, v \in B} w(u, v) \quad (2.5)$$

en este caso, la bipartición óptima del grafo será aquella que minimiza el valor de la función de corte. Sin embargo, este tipo de corte no presenta buenos resultados cuando se tienen nodos aislados en el grafo (ver Figura 2.7).

Por lo anterior, [Shi y Malik \(2000\)](#) proponen una variante a la función de corte. En lugar de fijarse en el valor total de las aristas que conectan las dos particiones, proponen una función que mida el corte como una fracción del total de conexiones de la partición hacia todos los nodos en el grafo. A esta medida la llamaron Normalized Cut (Corte Normalizado). Ambas funciones de corte puede ser vistas como medidas de disociación entre grupos o particiones.

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \quad (2.6)$$

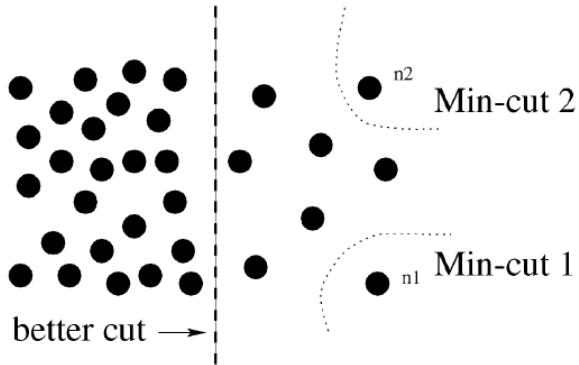


Figura 2.7: Caso donde el corte mínimo provee una mala partición.

Con esta nueva función de corte, los puntos aislados ya no tendrán un valor de corte pequeño, ya que el valor será con certeza un gran porcentaje de la conexión total de ese pequeño conjunto a todos los demás nodos. En la Figura 2.7 se observa que el valor *Min – cut1* del nodo n_1 será casi el 100 por ciento de la conexión total.

La desventaja de esta nueva función de corte es que su minimización es NP-completo, sin embargo, es posible aproximar una solución de manera eficiente a partir de eigenvalores y eigenvectores. Shi y Malik demostraron que la minimización de este criterio puede ser resuelto como un sistema de valores propios generalizados usando la matriz Laplaciana:

$$(D - W)\mathbf{y} = \lambda D\mathbf{y} \text{ sujeto a } \mathbf{y}^T D \mathbf{1} = 0$$

y demostraron que la restricción se auto satisface transformando el sistema a:

$$D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}\mathbf{z} = \lambda \mathbf{z} \text{ donde } \mathbf{z} = D^{\frac{1}{2}}\mathbf{y}$$

Es posible verificar que $D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$ es la matriz Laplaciana normalizada y que $z_0 = D^{\frac{1}{2}}\mathbf{1}$ corresponde al eigenvector de eigenvalor más pequeño (igual a 0) de dicha matriz. En esta aplicación de segmentación de imágenes, se utiliza el eigenvector de segundo eigenvalor más pequeño para partir el grafo en dos subconjuntos y cada subconjunto puede a su vez ser subdividido nuevamente hasta segmentar la imagen en las partes deseadas.

En resumen, los vectores propios son un medio por el cual se pueden identificar la particiones de un grafo, en este caso para fines de segmentación de una imagen.

Evaluación de características por Corte Normalizado

La evaluación de características por Corte Normalizado establece que dado un grafo \mathbb{G} y su matriz Laplaciana L , mediante la ecuación 2.7 se puede cuantificar cuanto varia localmente o que tan “suave” son los vectores de características $\mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n)^T \in \mathbb{R}^n$ sobre \mathbb{G} , entre más pequeño sea el valor de $\langle \mathbf{f}, L\mathbf{f} \rangle$ más suave será \mathbf{f} sobre \mathbb{G} .

$$\langle \mathbf{f}, L\mathbf{f} \rangle = \mathbf{f}^T L \mathbf{f} = \frac{1}{2} \sum_{v_i \sim v_j} w_{ij} (x_i - x_j)^2 \quad (2.7)$$

Un vector suave \mathbf{f} asigna valores similares a las instancias que están cerca entre sí en \mathbb{G} y por lo tanto será consistente con la estructura del grafo.

A partir de lo anterior, [Zhao y Liu. \(2007\)](#) proponen medir la consistencia de la i -ésima variable con la estructura del grafo mediante $\langle \mathbf{f}_i, L\mathbf{f}_i \rangle$, no obstante existen dos factores que afectan el valor que toma $\langle \mathbf{f}_i, L\mathbf{f}_i \rangle$, la norma de \mathbf{f}_i y L , los cuales deberán ser removidos ya que no contienen información sobre los datos pero que pueden causar que $\langle \mathbf{f}_i, L\mathbf{f}_i \rangle$ crezca o decrezca arbitrariamente. Estos dos factores pueden ser removidos mediante normalización de manera que:

$$\langle \mathbf{f}_i, L\mathbf{f}_i \rangle = \mathbf{f}_i^T L \mathbf{f}_i = \mathbf{f}_i^T D^{\frac{1}{2}} \mathcal{L} D^{\frac{1}{2}} \mathbf{f}_i = (D^{\frac{1}{2}} \mathbf{f}_i)^T \mathcal{L} (D^{\frac{1}{2}} \mathbf{f}_i) \quad (2.8)$$

Sea $\tilde{\mathbf{f}}_i = (D^{\frac{1}{2}} \mathbf{f}_i)$ el vector ponderado de la i -ésima característica y $\hat{\mathbf{f}}_i = \frac{\tilde{\mathbf{f}}_i}{\|\tilde{\mathbf{f}}_i\|}$ es el vector ponderado normalizado de la i -ésima característica.

Habiendo removido estos dos factores, el score de F_i (la i -ésima característica) se obtiene mediante la siguiente función:

$$\varphi_1(F_i) = \hat{\mathbf{f}}_i^T \mathcal{L} \hat{\mathbf{f}}_i = \frac{\mathbf{f}_i^T L \mathbf{f}_i}{\mathbf{f}_i^T D \mathbf{f}_i} \quad (2.9)$$

$\varphi_1(F_i)$ mide el valor del Corte Normalizado siendo \mathbf{f}_i el indicador de clúster para

particionar el grafo \mathbb{G} .

Rankeo de características mediante el Espectro del Grafo

Dada la matriz normalizada Lapaciana \mathcal{L} , se obtiene su descomposición espectral (λ_i, ξ_i) , donde λ_i es el eigenvalor y ξ_i el eigenvector de \mathcal{L} , con $0 \leq i \leq n - 1$, donde $\lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{n-1}$. De acuerdo a las propiedades de \mathcal{L} , $\lambda_0 = 0$ y $\xi_0 = D^{\frac{1}{2}}\mathbf{e}$, donde $\mathbf{e} = \{1, 1, \dots, 1\}^T$, usualmente (λ_0, ξ_0) es llamado el eigenpar trivial del grafo. Además, se puede demostrar que todos los eigenvalores de \mathcal{L} estarán contenidos en el intervalo $[0, 2]$.

Dada la descomposición espectral de \mathcal{L} es posible re escribir la ecuación 2.9 usando el eigensistema por medio del siguiente teorema:

Theorem 1. Sea $(\lambda_j, \xi_j), 0 \leq j \leq n - 1$ el eigensistema de \mathcal{L} , and $\alpha_j = \cos \theta_j$ donde θ_j es el ángulo entre \mathbf{f}_i and ξ_j . La ecuación 2.9 puede re-escribirse como:

$$\varphi_1(F_i) = \sum_{j=0}^{n-1} \alpha_j^2 \lambda_j, \quad \text{donde } \sum_{j=0}^{n-1} \alpha_j^2 = 1$$

Demostración. Sea $\Sigma = \text{DIAG}(\lambda_0, \lambda_1, \dots, \lambda_{n-1})$ y $U = (\xi_0, \xi_1, \dots, \xi_{n-1})$

Como $\|\widehat{\mathbf{f}}_i\| = \|\xi_j\| = 1$, entonces $\widehat{\mathbf{f}}_i^T \xi_j = \cos \theta_j$. Es posible re-escribir $\widehat{\mathbf{f}}_i^T \mathcal{L} \widehat{\mathbf{f}}_i$ como:

$$\widehat{\mathbf{f}}_i^T \mathcal{L} \widehat{\mathbf{f}}_i = \widehat{\mathbf{f}}_i^T U \Sigma U^T \widehat{\mathbf{f}}_i = (\alpha_0, \dots, \alpha_{n-1}) \Sigma (\alpha_0, \dots, \alpha_{n-1})^T = \sum_{i=0}^{n-1} \alpha_i^2 \lambda_i$$

Además, $\sum_{j=0}^{n-1} \alpha_j^2 = 1$, como $UU^T = I$ y $\|\widehat{\mathbf{f}}_i\| = 1$ □

El teorema anterior indica que el score de la i -ésima característica establecido en la ecuación 2.9 puede ser calculado como la combinación de eigenvalores de \mathcal{L} y $\cos \theta_1, \dots, \cos \theta_{n-1}$, estos últimos miden la similitud entre el vector de la i -ésima característica \mathbf{f}_i y los eigenvectores de \mathcal{L} . La teoría del *Clustering* Espectral señala que los eigenvalores de \mathcal{L} miden la separabilidad de los componentes o clústeres en un grafo y los eigenvectores fungen como indicadores de los clústeres. Eigenvalores iguales a cero indican el número de componentes independientes en un grafo, siempre

habrá al menos un componente, es decir, cuando todos los nodos están conectados y forma un sólo componente por esta razón $\lambda_0 = 0$. En la Figura 2.8 se puede apreciar el cambio en los valores de los eigenvalores de la matriz Laplaciana cuando tenemos 1 componente (todos los nodos conectados) y el caso donde hay 2 y 4 componentes independientes.

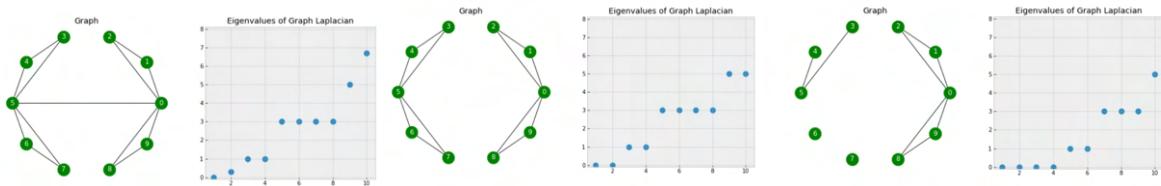


Figura 2.8: Grafos con 1, 2 y 4 componentes independientes [Fleshman \(2019\)](#).

Dado que $\lambda_0 = 0$, $\varphi_1(F_i)$ puede expresarse como

$$\varphi_1(F_i) = \hat{\mathbf{f}}_i^T \mathcal{L} \hat{\mathbf{f}}_i = \sum_{j=1}^{n-1} \alpha_j^2 \lambda_j \quad (2.10)$$

lo cual significa que el valor obtenido de la ecuación 2.9 mide la separabilidad del grafo usando a $\hat{\mathbf{f}}_i$ como el indicador de clústeres, dicha separabilidad es estimada al medir la similitudes entre $\hat{\mathbf{f}}_i$ y los vectores no triviales de \mathcal{L} .

Debido a que $\sum_{j=0}^{n-1} \alpha_j^2 = 1$ y $\alpha_0 \geq 0$, entonces $\sum_{j=1}^{n-1} \alpha_j^2 \leq 1$, lo que significa que entre mayor sea el valor de α_0^2 menor será el valor de $\sum_{j=1}^{n-1} \alpha_j^2$. Lo anterior implica que $\varphi_1(F_i)$ tome valores pequeños cuando $\hat{\mathbf{f}}_i$ es muy similar al eigenvector ξ_0 , sin embargo, dado que el eigenvector trivial ξ_0 sólo contiene información sobre densidad alrededor de la instancias y no determina la separabilidad, entonces será necesario realizar un ajuste que remueva este efecto. [Zhao y Liu. \(2007\)](#) proponen utilizar $\sum_{j=1}^{n-1} \alpha_j^2$ para normalizar $\varphi_1(F_i)$, lo cual resulta en la segunda función de rankeo:

$$\varphi_2(F_i) = \frac{\sum_{j=1}^{n-1} \alpha_j^2 \lambda_j}{\sum_{j=1}^{n-1} \alpha_j^2} = \frac{\hat{\mathbf{f}}_i^T \mathcal{L} \hat{\mathbf{f}}_i}{1 - \hat{\mathbf{f}}_i^T \xi_0} \quad (2.11)$$

Un valor pequeño en $\varphi_2(F_i)$ indica que $\hat{\mathbf{f}}_i$ se alinea estrechamente con los eigenvectores no triviales con autovalores pequeños y por lo tanto la i -ésima variable provee

una buena separabilidad de las observaciones. La teoría de *Clustering Espectral* indica que los primeros k eigenvectores de \mathcal{L} son los indicadores óptimos que determinan los clústeres que separan al grafo \mathbb{G} en k grupos. Por lo tanto, si es conocido el valor de k a priori, es decir, el caso supervisado, es posible usar la siguiente función de rankeo de características:

$$\varphi_3(F_i) = \sum_{j=1}^{k-1} (2 - \lambda_j) \alpha_j^2 \quad (2.12)$$

A diferencia de las dos funciones anteriores, $\varphi_3(F_i)$ asignará valores grandes a las variables que ofrecen mejor separabilidad debido a que un mayor score indica que la i -ésima característica se alinea estrechamente a los eigenvectores no triviales ξ_0, \dots, ξ_{k-1} , donde ξ_1 tiene la mayor prioridad. Al considerar sólo los k primeros eigenvectores $\varphi_3(F_i)$ logra un efecto de reducción de ruido.

Por último, existe una extensión de las funciones de rankeo $\varphi_1, \varphi_2, \varphi_3$ diseñadas para regularizar cuando se tienen variables que varían abruptamente en el grafo, en lugar de usar \mathcal{L} se utiliza $\gamma(\mathcal{L})$, donde $\gamma(\mathcal{L}) = \sum_{j=0}^{n-1} \gamma(\lambda_j) \xi_j \xi_j^T$ y $\gamma(\lambda_j)$ es una función creciente que penaliza, esta variante es de utilidad cuando el aprendizaje se realiza en datos muy ruidosos. No obstante, para fines del presente trabajo no se abordará en esta variante del algoritmo y por lo tanto supondremos que $\gamma(\mathcal{L}) = \mathcal{L}$

En resumen, el algoritmo SPEC seleccionará las variables más relevantes mediante la consistencia de la mismas con la estructura del grafo generado por S en tres pasos:

1. Se obtienen la matriz de similitud S y a partir de esta se construye el grafo de similitud \mathbb{G} y su correspondiente matriz de afinidad W .
2. Se obtiene la matriz Laplaciana normalizada \mathcal{L} y se evalúan las características usando el espectro de \mathcal{L} .
3. Se rankean los scores de las características en orden ascendente cuando se emplea la función $\widehat{\varphi}_1$ y $\widehat{\varphi}_2$, y en orden descendente cuando se emplea la función $\widehat{\varphi}_3$.

A continuación se muestra el algoritmo SPEC propuesto por [Zhao y Liu. \(2007\)](#): La complejidad del algoritmo recae únicamente en el costo computacional requerido para construir la matriz de similitudes y si fuese el caso el costo en calcular

Algorithm 1 SPEC

Input: $X, \gamma(\cdot), k, \hat{\varphi} \in \{\hat{\varphi}_1, \hat{\varphi}_2, \hat{\varphi}_3\}$

Output: SF_{SPEC}

▷ the ranked feature list

- 1: construct \mathbb{S} , the similarity set from X (and Y)
 - 2: construct graph \mathbb{G} from \mathbb{S} ;
 - 3: build W, D and \mathcal{L} from \mathbb{G} ;
 - 4: **for** each feature vector \mathbf{f}_i **do**
 - 5: $\widehat{\mathbf{f}}_i \leftarrow \frac{D^{\frac{1}{2}}\mathbf{f}_i}{\|D^{\frac{1}{2}}\mathbf{f}_i\|}; \quad SF_{SPEC}(i) \leftarrow \widehat{\varphi}(F_i)$
 - 6: ranking SF_{SPEC} in ascending order for $\widehat{\varphi}_1$ and $\widehat{\varphi}_2$, or descending order for $\widehat{\varphi}_3$;
 - 7: return SF_{SPEC} ;
-

$\gamma(\cdot)$. Por lo tanto, entre mayor sea el número de observaciones mayor será el costo computacional del algoritmo.

De acuerdo a lo anterior, resulta adecuado aplicar el algoritmo SPEC a los datos de este trabajo de tesis debido a que para cada estado se tienen a lo más 127 localidades por clusterizar.

El algoritmo SPEC puede aplicarse con distintas medidas de similitud, distintas funciones $\gamma(\cdot)$ y distintas funciones de rankeo $\varphi(\cdot)$, lo cual genera toda una familia de algoritmos de selección de características tanto para aprendizaje supervisado como no supervisado. En este trabajo de tesis se decide utilizar la función de similitud RBF kernel con una sutil pero importante modificación respecto a la propuesta presentada por [Zhao y Liu. \(2007\)](#), en lugar de elegir un sólo parámetro global σ se elegirá de manera local para cada observación un σ_i de manera que $s_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma_i \sigma_j}}$. Lo anterior se abordará con mayor detalle en el Capítulo 3.

Como se mencionó, para fines de este trabajo de tesis, se decide utilizar la función $\varphi_2(F_i)$ y la función $\gamma(\cdot) = \mathcal{L}$ de tal manera que el modelo con $\gamma(\cdot)$ no tenga efecto. En el capítulo 4 se presentan los resultados de aplicar el algoritmo SPEC a los datos de los 3 estados: Nayarit, Nuevo León y Yucatán.

Capítulo 3

Modelo

3.1. *Clustering* Espectral

La técnica de *clustering* o agrupación es una de las técnicas más utilizadas para el análisis exploratorio de datos, en prácticamente todos los campos científicos que tratan con datos empíricos, se intenta obtener una primera impresión de los datos tratando de identificar grupos de objetos con “comportamiento similar” con base en los valores de sus atributos [Luxburg \(2007\)](#).

En específico, el *Clustering* Espectral es una alternativa de agrupamiento que ha ganado popularidad en los últimos años debido a su facilidad de implementación, su capacidad de generar clústeres de alta calidad y de manejo de clústeres no convexos [Liu y Han \(2014\)](#). Es capaz de reconocer grupos cuando se tienen datos con formas complejas, figuras no lineales o figuras desconocidas [Zelnik-Manor y Perona \(2004\)](#), ya que no realiza supuestos sobre la forma de los clústeres y no requiere de estimar la distribución de los datos, por lo que muy a menudo supera a los algoritmos de *clustering* tradicionales como el algoritmo *K-means*. Algunos ejemplos de conjuntos de datos que son notoriamente más difíciles de clusterizar bajo algoritmos tradicionales se muestran a continuación:

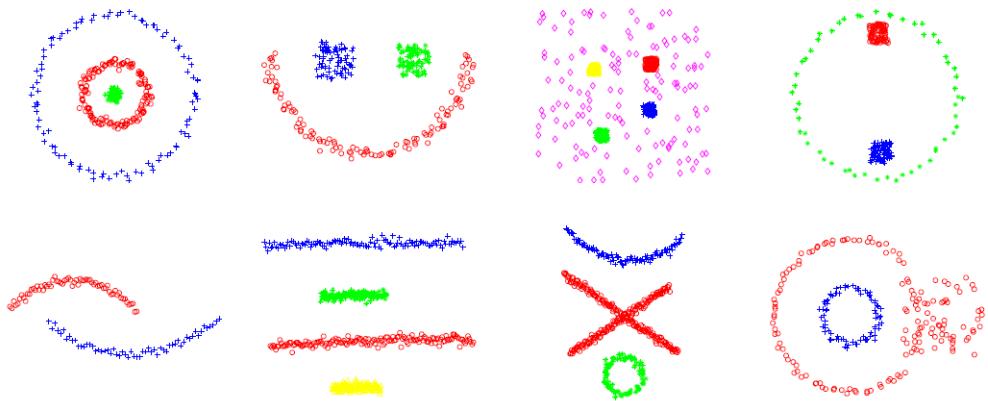


Figura 3.1: Conjuntos de datos con formas complejas Liu y Han (2014).

El *Clustering* Espectral provee buenos resultados cuando los grupos de observaciones forman regiones **no** convexas. Una región es convexa si cada par de puntos en la región puede ser unida por un segmento de recta y este segmento queda totalmente incluido en la región.

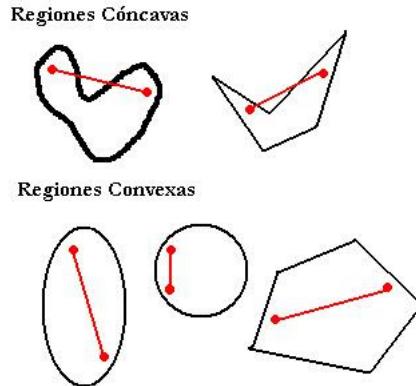


Figura 3.2: Regiones cóncavas y convexas.

La familia de algoritmos de *Clustering* Espectral se basa esencialmente en 3 pasos Liu y Han (2014):

1. Se construye el grafo de similitud de todas las instancias y su correspondiente matriz de adyacencia o afinidad.
2. Las instancias son transformadas a un espacio de menor dimensión (*embedding*), donde los clústeres son más obvios, con el uso de los eigenvectores de la matriz

Laplaciana. Esta representación en baja dimensión es conocida como *Spectral Embedding*, el cual también es usado como método de reducción de dimensiones.

3. Finalmente, se utiliza un algoritmo clásico como *K-means* para particionar las observaciones en el espacio del *embedding*.

En las siguientes figuras se puede observar un ejemplo de regiones no convexas donde el *clustering* mediante el algoritmo *K-means* resulta insatisfactorio, mientras que al proyectar los puntos al espacio de menor dimensión resulta bastante sencillo discriminar los grupos a través de cualquier algoritmo tradicional.

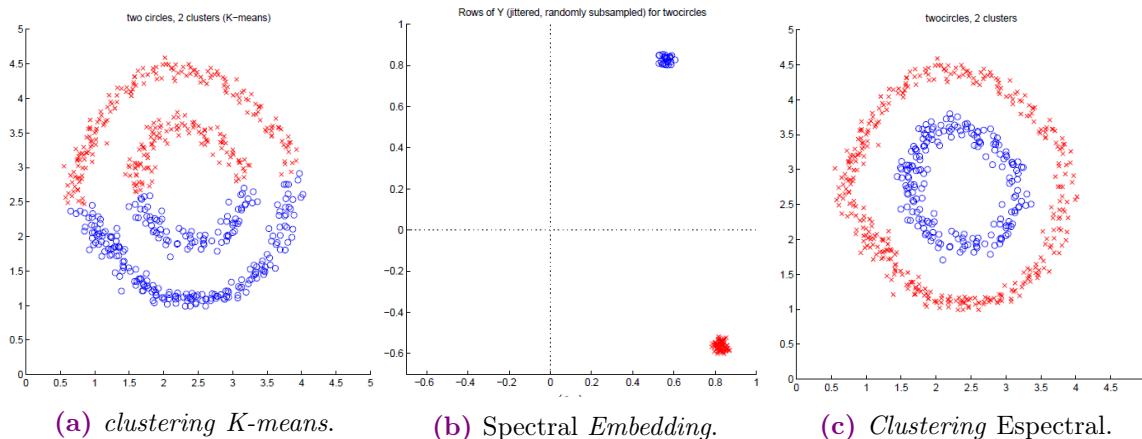


Figura 3.3: Ejemplo dona. *K-means* vs. *Clustering Espectral*

Los algoritmos de agrupamiento espectral trabajan con base en la matriz de similitud de los datos en lugar de utilizar los datos en su dimensión original. Esto implica un par de ventajas y desventajas, la mayor ventaja es poder trabajar con objetos que vienen de un espacio multidimensional y poder representarlos mediante nodos y aristas en un grafo. Por otro lado, la mayor desventaja es que puede llegar a ser computacionalmente costoso cuando el número de datos n es grande, ya que crear la matriz de similitud $n \times n$ y posteriormente sus vectores y valores propios puede implicar mucho tiempo y recurso computacional, por lo que no se recomienda cuando n es demasiado grande a menos que los datos sean extremadamente ruidosos y se encuentren en muy alta dimensión.

A continuación se desarrollan los conceptos base del *Clustering Espectral* para posteriormente detallar la metodología del algoritmo.

3.1.1. Grafo de Similitud

Dado un conjunto de n observaciones $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ en \mathbb{R}^m y sus similitudes s_{ij} entre todos los pares de instancias \mathbf{x}_i y \mathbf{x}_j , la mejor forma de representar los datos es mediante un grafo de similitud $\mathbb{G}(V, E)$. Cada instancia \mathbf{x}_i es representada mediante un vértice v_i y dos vértices v_i y v_j se conectan si la similitud entre ambas observaciones es mayor a cero o mayor a cierto umbral, los puentes o aristas que unen los vértices estarán determinados por s_{ij} . Bajo esta nueva representación de datos mediante el grafo de similitud, el problema de *clustering* puede ser re-formulado de la siguiente manera: se desea encontrar una partición del grafo tal que los puentes que conectan los nodos entre diferentes grupos tengan pesos (s_{ij}) muy bajos (lo que significa que las instancias entre diferentes grupos sean disímiles entre sí) y que los puentes que conectan los nodos dentro de un mismo grupo tengan pesos muy altos (lo que significa que instancias del mismo grupo sean muy similares entre ellas).

Existen diferentes grafos de similitud:

- Grafo ϵ -neighborhood: Dadas las distancias entre pares de instancias $d_{ij}^2 = \|\mathbf{x}_i - \mathbf{x}_j\|^2$, se conectarán los vértices cuya distancias sean menores a ϵ . Dado que todos los puentes que conectan las instancias serán menores a este umbral, incorporar el peso de los puentes que unen los vértices no aportará mayor información, por lo tanto se considera un grafo sin pesos.
- Grafo del k-vecino más cercano: Se conectará el vértice v_i con el vértice v_j si v_j está entre los k vecinos más cercanos de v_i o si v_i está entre los k vecinos más cercanos de v_j . La distancia para determinar los k vecinos usualmente es la norma l_1 , la norma l_2 o la distancia coseno. Dentro de este tipo de grafo existe una variante llamada “Grafo del k-vecino más cercano mútuo”, el cual conecta el vértice v_i con el vértice v_j si v_j está entre los k vecinos más cercanos de v_i y v_i está entre los k vecinos más cercanos de v_j . En ambos casos, después de conectar los vértices entre los k vecinos, se pueden asignar los pesos a los puentes que los unen mediante la similitud de sus puntos finales.
- Grafo totalmente conectado: Simplemente se conectan todos los vértices que

presenten una similitud positiva y a las aristas se les asigna el peso mediante s_{ij} . Ya que el grafo debería representar las relaciones de los vecinos locales, esta construcción de grafo solo es útil si la función de similitud en sí modela los vecindarios locales. Un ejemplo de tal función de similitud es la función de similitud Gaussiana o Radial Basis Function (RBF) Kernel:

$$s_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$$

, donde el parámetro σ controla el ancho de los vecindarios. Este parámetro juega un rol similar al parámetro ϵ en el caso del Grafo ϵ -neighborhood. En las siguientes secciones se revisará una propuesta para elegir el σ^2 adecuado de manera local para cada par de observaciones.

Como se puede observar, en los 3 tipos de grafos existe un parámetro a elegir por el usuario, el umbral ϵ , el número de vecinos k y el parámetro de escala σ^2 . El *Clustering* Espectral puede ser bastante sensible a cambios en los parámetros del grafo de similitud, por ello, la elección de dichos parámetros debe ser cuidadosa ya que el algoritmo puede variar considerablemente de acuerdo al valor elegido.

Los tipos de grafos mencionados son usados regularmente en el *Clustering* Espectral, sin embargo, de acuerdo a Luxburg (2007) no existe una respuesta concreta a la pregunta sobre qué tipo de grafo de similitud elegir. Como recomendación general Luxburg (2007) sugiere trabajar con grafos del k-vecino más cercano como primera opción, ya que de acuerdo a su experiencia es menos vulnerable a elecciones inadecuadas de parámetros. Asimismo, señala que no existe un estudio sistemático que investigue los efectos de elegir distintos grafos de similitud y sus parámetros (k , ϵ o σ^2), en su lugar se han propuesto varias recomendaciones pero sin un sustento teórico, por lo que dicha tarea se considera aún abierta para investigaciones futuras.

Por último, otro concepto importante a utilizar en el *Clustering* Espectral y que esta relacionado con el grafo de similitud es la matriz de adyacencia o afinidad W . Dadas las similitudes entre pares de observaciones s_{ij} , la matriz W indica qué nodos se encuentran conectados y está definida como $W(i, j) = s_{ij}$ para $i \neq j$ y $W(i, i) = 0$

para indicar que el nodo no está conectado consigo mismo.

3.1.2. Matriz Laplaciana de un grafo

La base del *Clustering* Espectral son las matrices Laplaciadas, el campo dedicado completamente al estudio de estas matrices es conocido como Teoría de Grafos Espectral. Existen diferentes tipos de matrices Laplaciadas, normalizadas y no normalizadas por lo tanto, la diferencia principal entre los algoritmos de *Clustering* Espectral radica en qué tipo de matriz Laplaciana se está utilizando.

Matriz Laplaciana No Normalizada: Sea D la matriz diagonal cuyo (i, i) -elemento es la suma de los elementos de la i -ésima fila de la matriz de adyacencia W (matriz de grado del grafo), la matriz Laplaciana no normalizada estará dada por $L = D - W$.

Algunas de las propiedades más importantes de L son:

1. Para cada vector vector de características $\mathbf{f} \in \mathcal{R}^n$

$$\mathbf{f}^T L \mathbf{f} = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (\mathbf{f}_i - \mathbf{f}_j)^2$$

2. L es simétrica y semidefinida positiva.
3. El eigenvalor más pequeño de L es 0 y su correspondiente eigenvector es el vector de unos $\mathbf{1}$.
4. L tiene n eigenvalores no negativos donde $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$

La matriz Laplaciana no normalizada, sus eigenvalores y eigenvectores pueden ser utilizados para describir algunas propiedades de los grafos. La siguiente propiedad es importante para llevar a cabo el *Clustering* Espectral:

- **Propiedad:** Sea \mathbb{G} un grafo no dirigido con pesos no negativos, la multiplicidad k del eigenvalor 0 de L es igual al número de componentes conectados A_1, A_2, \dots, A_k en el grafo. Los vectores propios de los eigenvalores 0 serán $\mathbf{1}_{A_1}, \mathbf{1}_{A_2}, \dots, \mathbf{1}_{A_K}$

Matriz Laplaciana Normalizada: Existen dos matrices Laplaciadas normalizadas:

$$\mathcal{L}_{sym} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

$$\mathcal{L}_{rw} = D^{-1} L = I - D^{-1} W$$

la primera es conocida como matriz simétrica Laplaciana y la segunda como matriz Laplaciana *random walk*.

Algunas de las propiedades más importantes de \mathcal{L}_{sym} y \mathcal{L}_{rw} son:

1. Para cada vector de características $\mathbf{f} \in \mathbb{R}^n$

$$\mathbf{f}^T \mathcal{L}_{sym} \mathbf{f} = \frac{1}{2} \sum_{i,j=1}^n w_{ij} \left(\frac{\mathbf{f}_i}{\sqrt{d_i}} - \frac{\mathbf{f}_j}{\sqrt{d_j}} \right)^2$$

2. λ es un eigenvalor de \mathcal{L}_{rw} con eigenvector \mathbf{u} si y sólo si λ es un eigenvalor de \mathcal{L}_{sym} con eigenvector $\mathbf{w} = D^{\frac{1}{2}} \mathbf{u}$.
3. λ es un eigenvalor de \mathcal{L}_{rw} con eigenvector \mathbf{u} si y sólo si λ y \mathbf{u} resuelve $L\mathbf{u} = \lambda D\mathbf{u}$.
4. 0 es un eigenvalor de \mathcal{L}_{rw} con eigenvector de unos $\mathbf{1}$. 0 es un eigenvalor de \mathcal{L}_{sym} con eigenvector $D^{\frac{1}{2}} \mathbf{1}$.
5. \mathcal{L}_{sym} y \mathcal{L}_{rw} son matrices semidefinidas positivas y tienen n eigenvalores no negativos donde $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$.

Las matrices Laplaciadas normalizadas, sus eigenvalores y eigenvectores pueden ser utilizados para describir algunas propiedades de los grafos. La siguiente propiedad es importante para llevar a cabo el *Clustering* Espectral:

- **Propiedad:** Sea \mathbb{G} un grafo no dirigido con pesos no negativos, la multiplicidad k del eigenvalor 0 de ambas matrices \mathcal{L}_{rw} y \mathcal{L}_{sym} es igual al número de componentes conectados A_1, A_2, \dots, A_k en el grafo. Los vectores propios de los eigenvalores 0 de \mathcal{L}_{rw} serán $\mathbf{1}_{A_1}, \mathbf{1}_{A_2}, \dots, \mathbf{1}_{A_K}$ y para \mathcal{L}_{sym} serán $D^{\frac{1}{2}} \mathbf{1}_{A_1}, D^{\frac{1}{2}} \mathbf{1}_{A_2}, \dots, D^{\frac{1}{2}} \mathbf{1}_{A_K}$

Una de las preguntas obligadas en el *Clustering* Espectral es ¿qué matriz Laplaciana es más conveniente utilizar?, de acuerdo con Luxburg (2007) se recomienda observar la distribución del grado del grafo de similitud \mathbb{G} . Si el grafo es muy regular y la mayoría de sus vértices tienen aproximadamente el mismo grado, entonces todas las matrices Laplaciadas serán muy similares entre sí y proporcionarán resultados similares en el *clustering*. Sin embargo, si los grados del grafo están distribuidos de forma heterogénea entonces las matrices Laplaciadas difieren considerablemente. Luxburg (2007) recomienda usar *Clustering* Espectral normalizado, y usar la matriz \mathcal{L}_{rw} en lugar de \mathcal{L}_{sym} . Un argumento a favor del *Clustering* Espectral normalizado, desde el punto de vista de la partición del grafo, es que el *Clustering* Espectral normalizado considera la minimización de las disimilitudes dentro del clúster y la maximización de disimilitudes entre los distintos clústeres, mientras que el *Clustering* Espectral no normalizado sólo considera la maximización de disimilitudes entre los distintos clústeres.

3.1.3. Algoritmos de *Clustering* Espectral

Como se mencionó anteriormente se han propuesto varios tipos de *Clustering* Espectral, sin embargo, la diferencia radica en el tipo de matriz Laplaciana que se utiliza para hacer la descomposición espectral:

1. *Clustering* Espectral no normalizado. Utiliza la matriz L
2. *Clustering* Espectral normalizado de acuerdo con Shi y Malik (2000). Utiliza la matriz \mathcal{L}_{rw} .
3. *Clustering* Espectral normalizado de acuerdo con Ng y cols. (2002). Utiliza la matriz \mathcal{L}_{sym} y realiza un paso adicional respecto a los otros 2 algoritmos, normaliza las filas de la matriz que guarda los eigenvectores de tal forma que tengan norma 1.

Debido a que los tres algoritmos son muy similares entre sí, a continuación se desarrolla el detalle sólo para el *Clustering* Espectral normalizado de acuerdo con Ng

y cols. (2002).

Siguiendo la notación de las secciones anteriores, dado un conjunto de n observaciones $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ en \mathbb{R}^m , las cuales se desean agrupar en k clústeres:

1. Se genera el grafo de similitud por alguna de las formas descritas en la sección 3.1.1 y se construye la matriz de afinidad o adyacencia $W \in \mathbb{R}^{n \times n}$. Ng y cols. (2002) sugieren utilizar un grafo totalmente conectado y la función de similitud RBF kernel, por lo tanto, W estará definida como $W_{ij} = s_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$ si $i \neq j$ y $W_{ii} = 0$.

$\|\mathbf{x}_i - \mathbf{x}_j\| = d_{ij}$ es la distancia euclíadiana entre dos puntos por lo que $W_{ij} = e^{-\frac{d_{ij}^2}{2\sigma^2}}$, donde σ^2 puede ser el parámetro de escala o de varianza de la función, cuya elección será discutida en la siguiente sección.

2. Se obtiene la matriz diagonal D cuyo (i, i) -elemento es la suma de la i -ésima fila de W y se construye la matriz Laplaciana normalizada $\mathcal{L}_{sym} = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$.
3. Se obtienen los primeros k eigenvectores (correspondientes a los k eigenvalores más pequeños). La representación de estos eigenvectores se denomina *Embedding Espectral* y funge como un método de reducción de dimensión.
4. Sea $U \in \mathbb{R}^{n \times k}$ la matriz que contiene los k eigenvectores como columnas, a partir de ella se construye la matriz $Z \in \mathbb{R}^{n \times k}$ normalizando las filas de U para que tengan norma 1, es decir, $z_{ij} = u_{ij}/(\sum_j u_{ij}^2)^{\frac{1}{2}}$
5. Considerando cada renglón de Z como una observación en \mathbb{R}^k , se agrupan en k clústeres mediante el algoritmo *K-means* o cualquier otro algoritmo buscando minimizar la distorsión.
6. Finalmente, se asigna a la instancia original \mathbf{x}_i el clúster j si y sólo si la fila i de la matriz Z ha sido asignado al clúster j .

3.1.4. Función de similitud RBF kernel y parámetro de escala local σ^2

La función de similitud Radial Basis Function kernel, RBF kernel o kernel Gaussiano es un kernel que tiene la forma de la función Gaussiana:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$$

donde σ determina el ancho del kernel Gaussiano (o su equivalente $\gamma = \frac{1}{2\sigma^2}$) y $\|\mathbf{x}_i - \mathbf{x}_j\| = d_{ij}$ es la distancia euclíadiana entre los dos puntos. Es considerada función de similitud ya que su valor se encontrará entre 0 y 1, cuando $d_{ij} = 0$ la función es igual a 1, indicando que las dos observaciones son iguales. Por el contrario, cuando d_{ij} es extremadamente grande el valor de la función tiende a 0.

Un kernel es cualquier función de la forma:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \langle \psi(\mathbf{x}_1), \psi(\mathbf{x}_2) \rangle$$

donde ψ es una función que proyecta al vector \mathbf{x}_i a un nuevo espacio vectorial y la función kernel K calcula el producto punto entre los dos vectores proyectados.

Como se prueba a continuación, la función ψ en un RBF kernel proyecta los vectores a un espacio de infinitas dimensiones **Bernstein** (s.f.).

Demostración:

Sin pérdida de generalidad, supongamos $\sigma = 1$

$$\begin{aligned} K_{RBF}(\mathbf{x}_1, \mathbf{x}_2) &= \exp \left[-\frac{1}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \right] \\ &= \exp \left[-\frac{1}{2} \langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j \rangle \right] \\ &= \exp \left[-\frac{1}{2} (\langle \mathbf{x}_i, \mathbf{x}_i - \mathbf{x}_j \rangle - \langle \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j \rangle) \right] \\ &= \exp \left[-\frac{1}{2} (\langle \mathbf{x}_i, \mathbf{x}_i \rangle - \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \langle \mathbf{x}_j, \mathbf{x}_i \rangle + \langle \mathbf{x}_j, \mathbf{x}_j \rangle) \right] \\ &= \exp \left[-\frac{1}{2} (\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\langle \mathbf{x}_i, \mathbf{x}_j \rangle) \right] \end{aligned}$$

$$\begin{aligned}
K_{RBF}(\mathbf{x}_i, \mathbf{x}_j) &= \underbrace{\exp \left[-\frac{1}{2} \|\mathbf{x}_i\|^2 - \frac{1}{2} \|\mathbf{x}_j\|^2 \right]}_C \exp [\langle \mathbf{x}_i, \mathbf{x}_j \rangle] \\
&= C e^{\langle \mathbf{x}_i, \mathbf{x}_j \rangle} \\
&= C \sum_{n=0}^{\infty} \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle^n}{n!} \text{ Expansión de Taylor de } e^x \\
&= C \sum_{n=0}^{\infty} \frac{K_{poly(n)}(\mathbf{x}_i, \mathbf{x}_j)}{n!}
\end{aligned}$$

En la demostración anterior es posible notar que el kernel RBF se compone de la suma infinita de kernels polinomiales de grado n . Por lo tanto, esta función de similitud es capaz de medir relaciones no lineales entre dos observaciones en infinitas dimensiones.

A manera de ejemplo y para dar una intuición de como funciona el kernel Gaussiano, en la Figura 3.4 se muestra la aplicación de un kernel polinomial de grado 2, el cual mapea puntos de \mathbb{R}^2 a \mathbb{R}^3 , en esta nueva proyección los grupos son más obvios y las observaciones son linealmente separables.

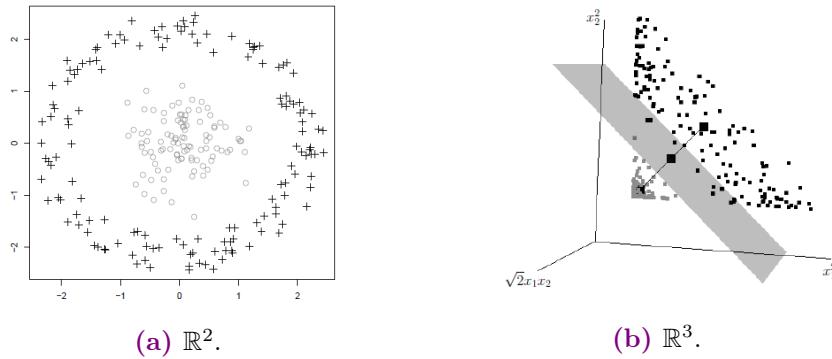


Figura 3.4: Kernel polinomial de grado 2 sobre datos en \mathbb{R}^2

Un elemento de suma importancia a elegir es el valor del σ^2 . Este parámetro controla qué tan rápidamente la matriz W decae conforme la distancia euclíadiana entre \mathbf{x}_i y \mathbf{x}_j aumenta, en cierta forma el parámetro marca el umbral o zona donde dos puntos son considerados similares o no. En la Figura 3.5 se muestra el efecto que

tiene σ^2 en la función de similitud RBF kernel, valores muy pequeños de σ^2 generan vecindades muy reducidas ya que sólo las observaciones muy cercanas se considerarán similares. Por el contrario, cuando σ^2 es grande, la vecindad es mucho más amplia.

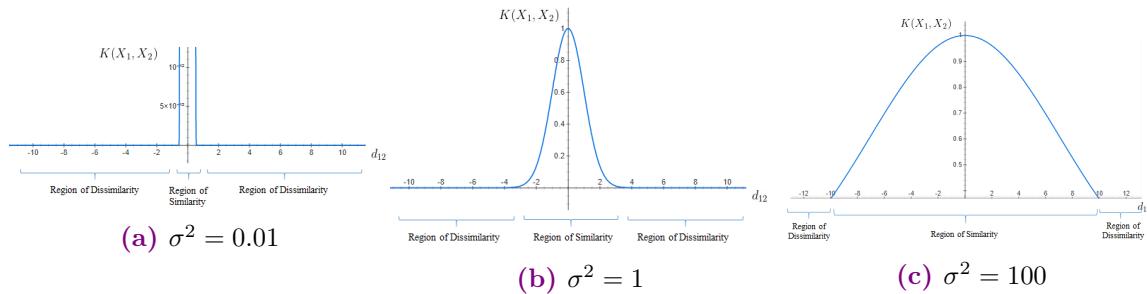


Figura 3.5: Función RBF bajo distintos parámetros de σ^2 Sreenivasa (2020).

En este punto, la pregunta es ¿cómo elegir el parámetro σ^2 adecuado?. Ng y cols. (2002) proponen seleccionar σ^2 mediante la repetición del algoritmo de *Clustering* Espectral para diferentes valores y elegir el σ^2 que provea los clústeres menos distorsionados (con mejor ajuste). Sin embargo, este proceso conduce a un mayor costo computacional y persiste la parte manual de proponer distintos valores de σ^2 .

Además, cuando los datos incluyen clústeres con diferentes estadísticos locales puede ser que no exista un sólo valor de σ^2 eficiente para todo el conjunto de datos Zelnik-Manor y Perona (2004). Para ilustrar el efecto que tiene la elección de σ^2 en el *Clustering* Espectral se muestra la Figura 3.6. Se puede observar como ligeros cambios en el valor de σ^2 afectan significativamente los clústeres generados debido a que los datos contienen distintas dispersiones entre los grupos.

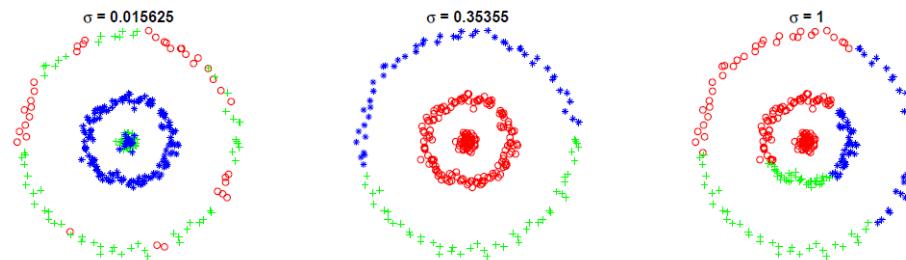


Figura 3.6: *Clustering* Espectral sin parámetro de escala local Zelnik-Manor y Perona (2004).

Por lo anterior, Zelnik-Manor y Perona (2004) proponen usar un parámetro de escala local σ_i para cada observación \mathbf{x}_i en lugar de un solo valor σ para todos los datos.

La distancia de \mathbf{x}_i a \mathbf{x}_j respecto a \mathbf{x}_i esta dada por $\frac{d(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_i}$ mientras que la distancia respecto \mathbf{x}_j esta dada por $\frac{d(\mathbf{x}_j, \mathbf{x}_i)}{\sigma_j}$, por tanto, la distancia d^2 puede ser generalizada como $\frac{d(\mathbf{x}_i, \mathbf{x}_j)d(\mathbf{x}_j, \mathbf{x}_i)}{\sigma_i\sigma_j} = \frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{\sigma_i\sigma_j}$.

Considerando esta propuesta, la matriz de afinidad queda definida por $\hat{W}_{ij} = e^{-\frac{d^2(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_i\sigma_j}}$ para $i \neq j$ y $W_{ii} = 0$ cuando $i = j$, donde $d(\mathbf{x}_i, \mathbf{x}_j)$ es alguna función de distancia, comúnmente la distancia Euclidiana.

Usando la expresión anterior, el parámetro de escala se auto ajustará de acuerdo a las estadísticas locales de los vecinos alrededor de \mathbf{x}_i y \mathbf{x}_j . En la Figura 3.7 se puede apreciar el efecto de utilizar un parámetro de escala local, se tienen dos grupos uno con menor dispersión que el otro, la afinidad entre cada punto y sus vecinos circundantes se indica por el grosor de la línea que los conecta. En la segunda figura, se aplica un mismo parámetro σ^2 para todos los datos, y vemos que las afinidades entre los clústeres son más grandes que las afinidades dentro de cada grupo. La tercera figura corresponde a las afinidades al haber aplicado el parámetro local, en este caso las afinidades entre los clústeres ahora son significativamente menores que las afinidades dentro de los grupos.

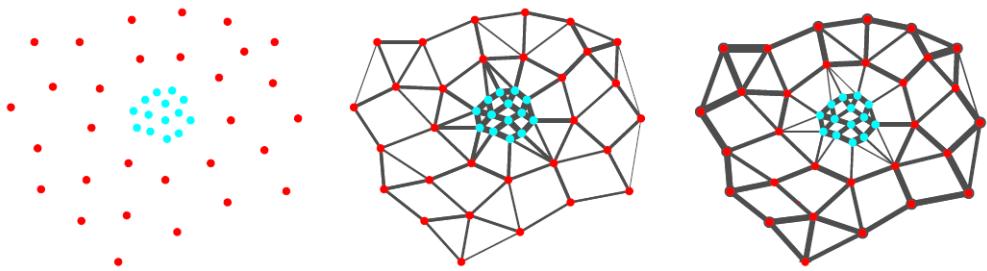


Figura 3.7: Efecto del parámetro de escala local [Zelnik-Manor y Perona \(2004\)](#).

La selección del parámetro de escala local σ_i puede ser determinado mediante el estudio de las estadísticas locales de los vecinos de la observación \mathbf{x}_i . La propuesta que realizan los autores es:

$$\sigma_i = d(\mathbf{x}_i, \mathbf{x}_K) \quad (3.1)$$

donde \mathbf{x}_K es el K -ésimo vecino de \mathbf{x}_i . La selección de K esta en función de la dimensión de espacio de *embedding*, sin embargo, [Zelnik-Manor y Perona \(2004\)](#) men-

cionan que con $K = 7$ han obtenido buenos resultados en diferentes experimentos con datos de alta dimensionalidad. No obstante, cabe señalar que este método de selección de parámetros conlleva un mayor costo computacional ya que calcula el K vecino más cercano para cada observación.

Para fines del presente trabajo se decide utilizar esta propuesta de parámetro de escala local en el algoritmo de *Clustering Espectral* propuesto por Ng y cols. (2002).

3.1.5. Número de clústeres

Como se revisó en la sección 3.1.3, la fase final del algoritmo de *Clustering Espectral* propuesto por Ng y cols. (2002) incluye la aplicación del algoritmo K -means sobre el espacio de dimensión reducida conocido como *Embedding Espectral*.

En ese contexto, uno de los factores de mayor impacto en este tipo de *clustering* es la elección del número correcto de clústeres. Diversas investigaciones han propuesto distintos criterios para la elección del valor óptimo de K , algunos criterios que han destacado y que no asumen ningún supuesto distribucional en los datos son: Índice Calinski-Harabasz, Coeficiente Silhouette e Índice Davies Bouldin Aggarwal y Reddy (2014).

Índice Calinski-Harabasz

El Índice Calinski-Harabasz se define de la siguiente manera:

$$CH(K) = \frac{\frac{B(K)}{K-1}}{\frac{W(K)}{N-K}} \quad (3.2)$$

El número óptimo de grupos K es el valor que maximiza la ecuación 3.2, N representa el número de instancias, $B(K)$ representa la dispersión entre los centroides de cada clúster y el centroide total de los datos multiplicado por el número de instancias en cada clúster (between clusters sum of squares BCSS) y $W(K)$ representa la dispersión dentro de cada clúster, entre los puntos y su correspondiente centroide (within-cluster

sum of squares WCSS). Dichos términos se calculan de la siguiente forma:

$$W(K) = \sum_{q=1}^K \sum_{\mathbf{x}_i \in C_q} \|\mathbf{x}_i - \bar{x}_q\|^2 = \sum_{q=1}^K \sum_{\mathbf{x}_i \in C_q} d(\mathbf{x}_i, \bar{x}_q)^2$$

$$B(K) = \sum_{q=1}^K N_q \|\bar{x}_q - \bar{x}\|^2 = \sum_{q=1}^K N_q d(\bar{x}_q, \bar{x})^2$$

, C_q es el conjunto de puntos en el clúster q , \bar{x}_q centroide del clúster q , N_q número de elementos en el clúster q y \bar{x} es el centroide del total de observaciones.

Coeficiente Silhouette

Este criterio considera tanto la dispersión intra-clúster como entre-clústeres.

Dada una instancia \mathbf{x}_i , su coeficiente Silhouette $s(i)$ se calcula por medio de la siguiente ecuación:

$$s(i) = \frac{b(i) - a(i)}{\max(a_i, b_i)} \quad (3.3)$$

Primero se obtiene $a(i)$, la distancia promedio de \mathbf{x}_i a todos los puntos de su mismo clúster. Después, se calcula el promedio de las distancias entre \mathbf{x}_i y todos los puntos del clúster más cercano, que corresponde al valor de $b(i)$. De manera que el coeficiente $s(i)$ mide la diferencia estandarizada entre $b(i)$ y $a(i)$, por lo que $s(i) \in [-1, 1]$. Cuando $s(i)$ se acerca a 1, la instancia \mathbf{x}_i es más cercana a su propio clúster y se considera bien clasificado. Cuando $s(i)$ se acerca a -1 se considera más cercano a otro clúster y por lo tanto se considera una mala clasificación.

Los valores que puede tomar $s(i)$ se pueden expresar de la siguiente manera:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{si } a(i) < b(i) \\ 0, & \text{si } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & \text{si } a(i) > b(i) \end{cases} \quad (3.4)$$

Por último, se obtiene el promedio de todos los coeficientes Silhouettes en el conjunto de datos y se elige el valor de K que maximize el coeficiente S Rousseeuw

(1987):

$$S = \frac{\sum_{i=1}^N \frac{b(i)-a(i)}{\max(a_i, b_i)}}{N} \quad (3.5)$$

Índice Davies Bouldin

Este índice calcula la similitud promedio entre cada clúster C_i con $i = 1, 2, \dots, K$ y su clúster más cercano C_j . La similitud es una medida que compara la distancia dentro de los grupos con la distancia de los grupos en sí. Un valor bajo en el índice indica mejor separabilidad entre los clústeres. El valor más pequeño que puede tomar el índice es 0.

La similitud está definida como la medida R_{ij} definida por:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (3.6)$$

donde s_i es la distancia promedio entre cada punto del clúster i y el centroide de ese clúster, y d_{ij} es la distancia entre el centroide del clúster i y clúster j .

Por último, el índice se calcula como el promedio de las máximas similitudes:

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} R_{ij} \quad (3.7)$$

La expresión $\max_{j \neq i} R_{ij}$ indica que sólo se considera el valor R_{ij} del clúster más similar C_j para cada C_i Halkidi, Batistakis, y Vazirgiannis (2001).

Eigengap y Búsqueda iterativa del Eigengap

Los criterios anteriores pueden ser usados en el *Clustering* Espectral, aunque de manera adicional, existe una herramienta diseñada en particular para este tipo de *clustering*, se denomina heurístico Eigengap y mide la estabilidad de los eigenvectores de la matriz Laplaciana, ya sea de tipo normalizada o no normalizada.

El heurístico Eigengap tiene como objetivo elegir el número de clústeres k tal que los eigenvalores $\lambda_1, \dots, \lambda_k$ son muy pequeños, pero λ_{k+1} es relativamente grande. Parte de la justificación del uso de este heurístico se encuentra basada en la teoría

de perturbación de matrices donde el subespacio abarcado por los primeros k autovectores de la matriz Laplaciana es estable si y solo si el eigengap $\delta_k = |\lambda_k - \lambda_{k+1}|$ es grande. Siendo n el número de observaciones, $\delta_1, \delta_2, \dots, \delta_{n-1}$ son los eigengaps y $\lambda_1, \lambda_2, \dots, \lambda_n$ son los eigenvalores de la matriz Laplaciana.

El heurístico eigengap suele funcionar bien en conjuntos bien separados. Esto se puede observar en el caso donde se tienen k clústeres completamente desconectados en donde el eigenvalor 0 tiene multiplicidad k y posteriormente hay un gap o brecha con el eigenvalor $\lambda_{k+1} > 0$.

A continuación se muestra un ejemplo de cómo funciona el heurístico eigengap cuando se tienen grupos bien definidos y cuando se tienen grupos traslapados.

Se generaron muestras de datos en \mathbb{R} de cuatro distribuciones Gaussianas con varianza creciente, podemos observar que el histograma de la izquierda presenta menor varianza y por lo tanto clústeres más definidos, y el histograma de la derecha tiene mayor varianza y por lo tanto clústeres traslapados. Con base en estos datos se construyó un grafo de similitud 10 vecinos más cercanos y se grafican los eigenvalores obtenidos de la matriz Laplaciana normalizada \mathcal{L}_{rw} (puede utilizarse cualquier matriz Laplaciana de las anteriormente citadas).

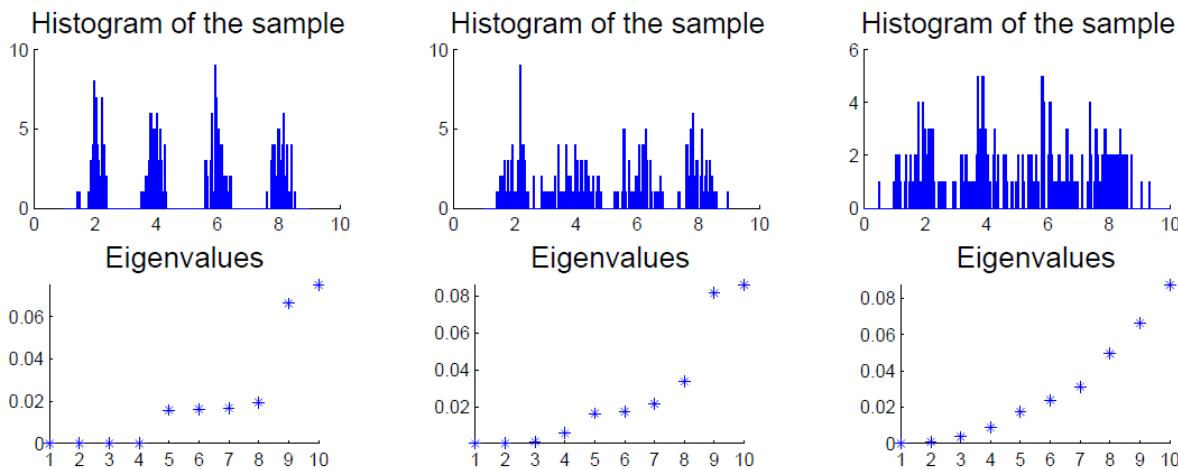


Figura 3.8: Tres conjuntos de datos y los 10 eigenvalores más pequeños de \mathcal{L}_{rw} Luxburg (2007).

El primer conjunto de datos tiene los 4 clústeres bien definidos y se puede observar que los primeros 4 eigenvalores son aproximadamente 0, posteriormente se observa un

gap entre el cuarto y el quinto eigenvalor, $|\lambda_5 - \lambda_4|$ es relativamente grande. De acuerdo con el estadístico eigengap, los datos presentan 4 clústeres. El segundo conjunto se comporta de forma similar al anterior, sin embargo, empieza a ser poco notoria la diferencia entre eigenvalores. En el tercer conjunto de datos, no se observa un gap bien definido ya que parece que la diferencia entre todos los eigenvalores es la misma y esto se debe a que los clústeres se traslanan, incluso para el ojo humano resulta difícil distinguir los grupos. Lo anterior demuestra que, al igual que muchos otros criterios de selección del valor óptimo de k , el heurístico eigengap usualmente trabaja bien si los datos contienen grupos bien definidos pero ofrece resultados ambiguos cuando se tienen grupos traslapados.

Adicionalmente, [Afzalan y Jazizadeh \(2019\)](#) proponen una nueva variante para el heurístico eigengap denominada Búsqueda iterativa del Eigengap. La propuesta surge en el contexto de conjuntos de datos con diferentes escalas.

Esta metodología va descubriendo agrupaciones mediante una búsqueda iterativa a largo de una estructura en forma de árbol. En cada iteración puede ser que el heurístico eigengap no haya encontrado los clústeres finales, la idea es que el algoritmo vaya refinando los clústeres en cada iteración ya que visitará cada clúster resultante de la iteración anterior y calculará nuevamente el heurístico eigengap para determinar si existen nuevos clústeres dentro del clúster visitado. El criterio de parada será cuando se hayan visitado todos los nodos del árbol y en todos los nodos finales (leaf nodes) el eigengap indique que sólo existe un clúster.

En la siguiente Figura se ilustra el proceso de Búsqueda iterativa del Eigengap. En la estructura del árbol, la raíz indica el conjunto completo de observaciones al cual se aplica el heurístico Eigengap y el *Clustering* Espectral para determinar el número inicial de clústeres K . Cada uno de los nodos hijos serán los K clústeres generados de la iteración anterior, a los cuales se les aplicará nuevamente el heurístico Eigengap y el *Clustering* Espectral para determinar si existen nuevos clústeres los cuales a su vez serán los nuevos nodos hijos. La búsqueda terminará cuando el heurístico eigengap indique que el número de clústeres es $K = 1$ en todos los nodos finales.

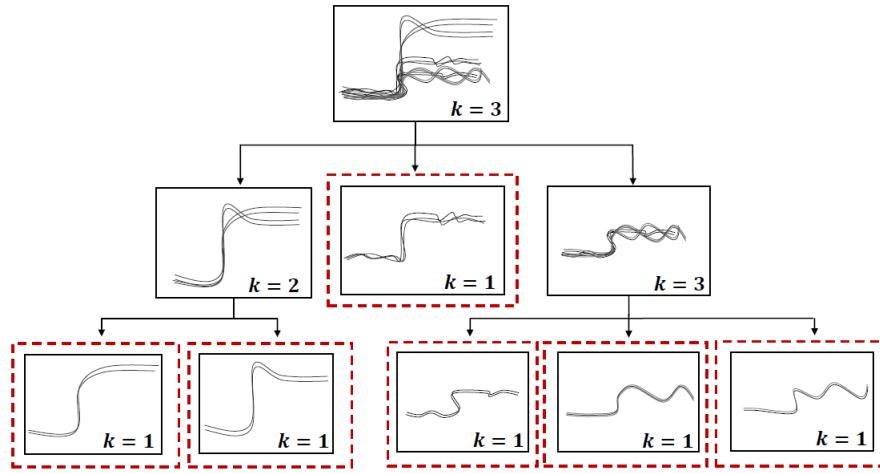


Figura 3.9: Diagrama Búsqueda Iterativa del Eigengap.

En la siguiente sección se muestran los resultados obtenidos al aplicar los modelos desarrollados en este capítulo sobre datos reales con información de estado de Nayarit, Nuevo León y Yucatán.

Capítulo 4

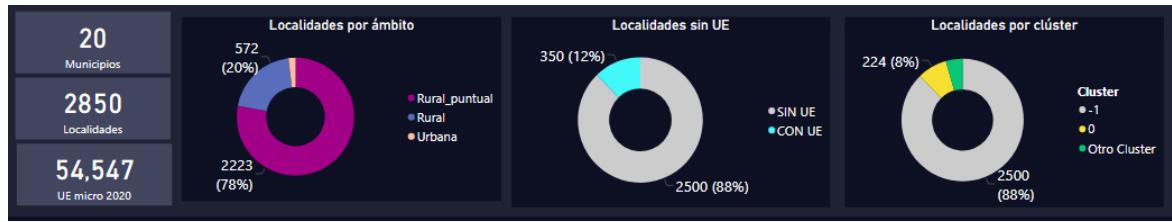
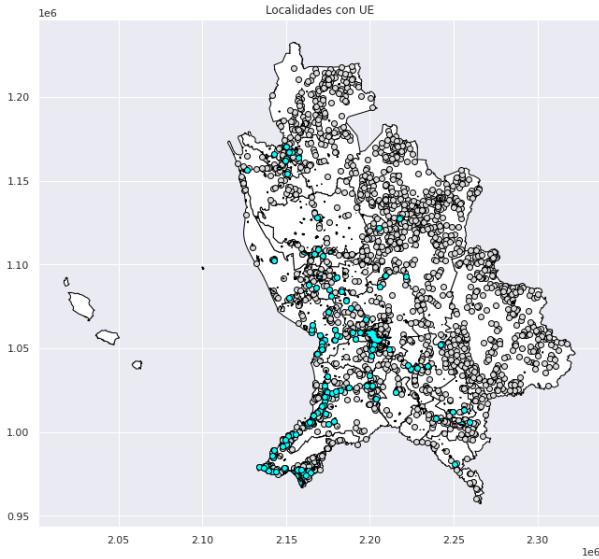
Modelación

A lo largo de este capítulo se presenta el resultado final de aplicar las metodologías estudiadas en los capítulos anteriores. Su contenido está dividido en tres secciones, cada sección corresponde a un estado, se inicia con un análisis exploratorio de los datos, posteriormente se muestran los resultados de la técnica de selección de variables SPEC y del algoritmo de *clustering* espectral. Por último, se realiza una interpretación de los resultados por clúster y finalmente se muestra el ranking de localidades por medio de mapas.

4.1. Nayarit

4.1.1. Análisis exploratorio

El estado de Nayarit cuenta con 20 municipios y 2850 localidades con más de 1 o 2 viviendas, la mayoría de estas localidades son de tipo rural puntual (78 %), esto quiere decir que son localidades rurales no amanzanadas. Asimismo, cabe resaltar que 87.7 % de las localidades han registrado 0 Unidades Económicas (UE) micro en el DENUE, por lo que es conveniente separar dichas localidades del resto dentro de un *Cluster -1* que representará nula actividad del sector micro. En la Figura 4.2 se puede observar en color azul las 350 localidades que si han registrado unidades económicas, las cuales en su mayoría se encuentran al suroeste del estado.

**Figura 4.1:** Información general Nayarit**Figura 4.2:** Localidades con Unidades Económicas

De las 350 localidades que si cuentan con UE, 141 corresponden a localidades rurales puntuales donde la mayoría de casos registran menos de 5 UE (Figura 4.3), es decir, son localidades de 6 habitantes en promedio que presentan poca o casi nula actividad económica en el sector micro, por tanto conviene agruparlos desde un inicio en un *Cluster 0* de muy baja actividad económica micro.

De manera similar, 154 son localidades rurales amanzanadas y aunque éstas tienen un mayor número de habitantes (alrededor de 700), una gran cantidad tiene menos de 5 UE, y por lo tanto también se incluyen al *Cluster 0*. En el caso de las localidades urbanas, sólo una localidad tiene menos de 5 UE, la cual se incluye en el *Cluster 0*.

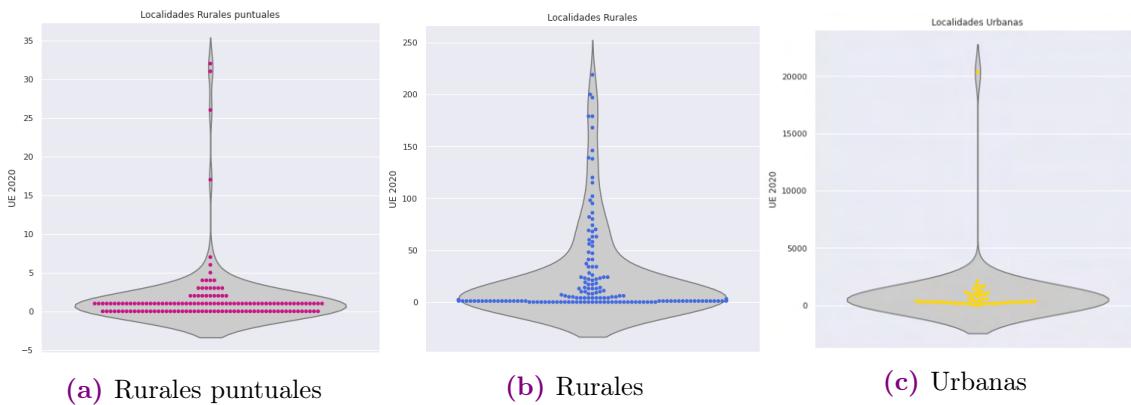


Figura 4.3: UE 2020. Localidades por ámbito.

Considerando lo anterior, de las 2850 localidades con más de 1 o 2 viviendas, 2500 pertenecen al *Cluster -1* de nula actividad económica micro y 224 pertenecen al *Cluster 0* de poca actividad económica Micro y las 126 restantes se clusterizarán a partir de las variables referentes a la actividad económica micro obtenidas del DENUE y las variables socio-demográficas obtenidas del CPV.

Cluster	Descripción	Localidades	Urbanas	Rurales	Rurales puntuales
-1	Nula actividad micro	2500	0	418	2082
0	Baja actividad micro	224	1	88	135
Otros	Loc. por clusterizar	126	54	66	6

Tabla 4.1: Distribución localidades Nayarit. *Cluster -1*, *Cluster 0* y Otros.

A continuación se muestran las localidades que quedaron agrupadas en el *Cluster -1* (gris), *Cluster 0* (amarillo) y Otros(verde).

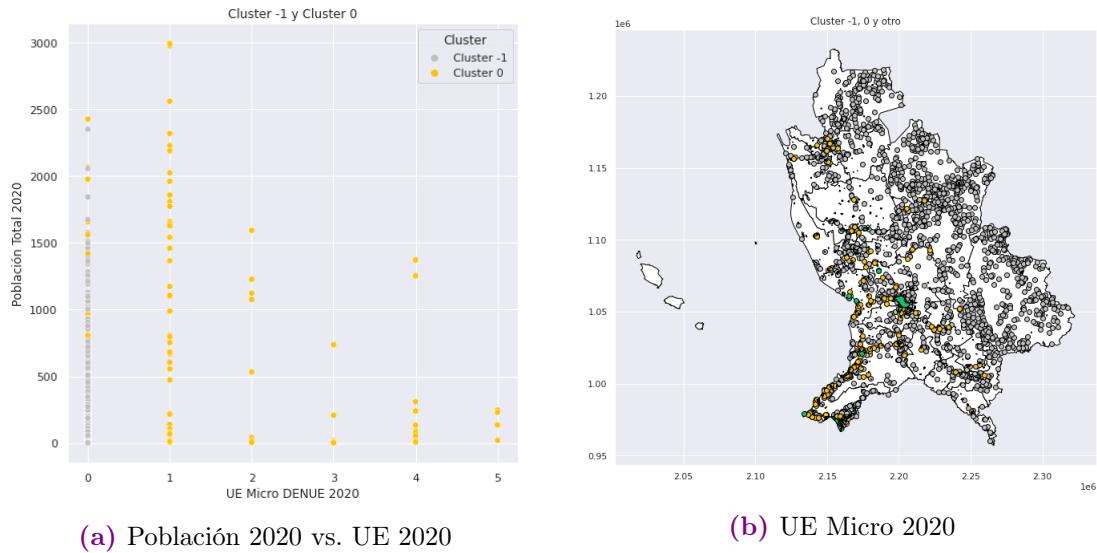


Figura 4.4: Distribución localidades Nayarit. Cluster -1, 0 y Otros.

Como se puede apreciar en la Figura 4.4 hay casos en los que las localidades cuentan con un mayor número de habitantes pero sólo tienen registro de 0 o 1 UE, esto apunta a que el alcance de las actualizaciones del DENUE es limitado ya constantemente se están generando nuevos negocios, por lo que llevar a cabo su actualización total debería de ser con una periodicidad menor a la actual que es cada 5 años.

4.1.2. Selección de variables mediante algoritmo SPEC.

Considerando la base de datos de las 126 localidades a clusterizar, se aplica el algoritmo SPEC a 3 conjuntos de variables: 1) Variables económicas del DENUE, 2) Variables socio-demográficas del CPV y 3) Variables del DENUE y del CPV. Se aplicará de esta manera para poder identificar de manera individual y en conjunto cuáles son las variables que tienen mayor influencia en la formación de los clústeres.

Como se mencionó en la sección 2.3.1, para fines de este trabajo se decide utilizar la función de similaridad RBF kernel propuesta por [Zhao y Liu. \(2007\)](#) y en lugar de elegir un sólo parámetro global σ se opta por asignar dicho parámetro de manera local, por lo que $s_{ij} = e^{-\frac{\|\mathbf{x}_i \mathbf{x}_j\|^2}{2\sigma_i \sigma_j}}$, donde $\sigma_i = d(\mathbf{x}_i, \mathbf{x}_K)$ con \mathbf{x}_K igual al séptimo vecino de \mathbf{x}_i tal como lo propone [Zelnik-Manor y Perona \(2004\)](#).

Además, se decide utilizar la función $\varphi_2(F_i)$ y la función $\gamma(\cdot) = \mathcal{L}$ de tal manera que el modelo con $\gamma(\cdot)$ no tenga efecto. Por lo tanto, valores pequeños en el score $\varphi_2(F_i)$ indican que la variable F_i se alinea estrechamente con los eigen-vectores no triviales de los eigenvalores más pequeños y por lo tanto la i -ésima variable provee una buena separabilidad en las observaciones.

El primer conjunto de variables a las que se le aplicará el algoritmo SPEC son las variables económicas del DENUE. Siguiendo los pasos del algoritmo SPEC, se obtiene la matriz de similaridad S , con base en esta se construye el grafo totalmente conectado no dirigido cuyos pesos están determinados por las s_{ij} y su correspondiente matriz de afinidad W . En las Figuras 4.6a y 4.6b se puede observar el mapa de calor de W y la representación del grafo completamente conectado, respectivamente. Asimismo, en la Figura 4.6c se muestra el grafo de los 10-vecinos más cercanos con fines ilustrativos, ya que este tipo de grafo permite visualizar los grupos de manera más clara.

Posteriormente, se obtiene la matriz Laplaciana normalizada \mathcal{L} y su descomposición espectral, con base en ella se obtendrán los scores $\varphi_2(F_i)$ para cada variable. En la Figura 4.6e se puede notar que las variables que más influyen en la clusterización son las referentes a la tasa de sobre-vivencia de las UE micro por antigüedad y las de menor relevancia son las variables que indican la proporción de UE micro que incre-

mentaron de personal y que por lo tanto pasaron de ser micro a medianas y grandes empresas, esto tiene bastante sentido ya que son muy pocas las observaciones que registran saltos tan grandes en cuanto a número de personal.

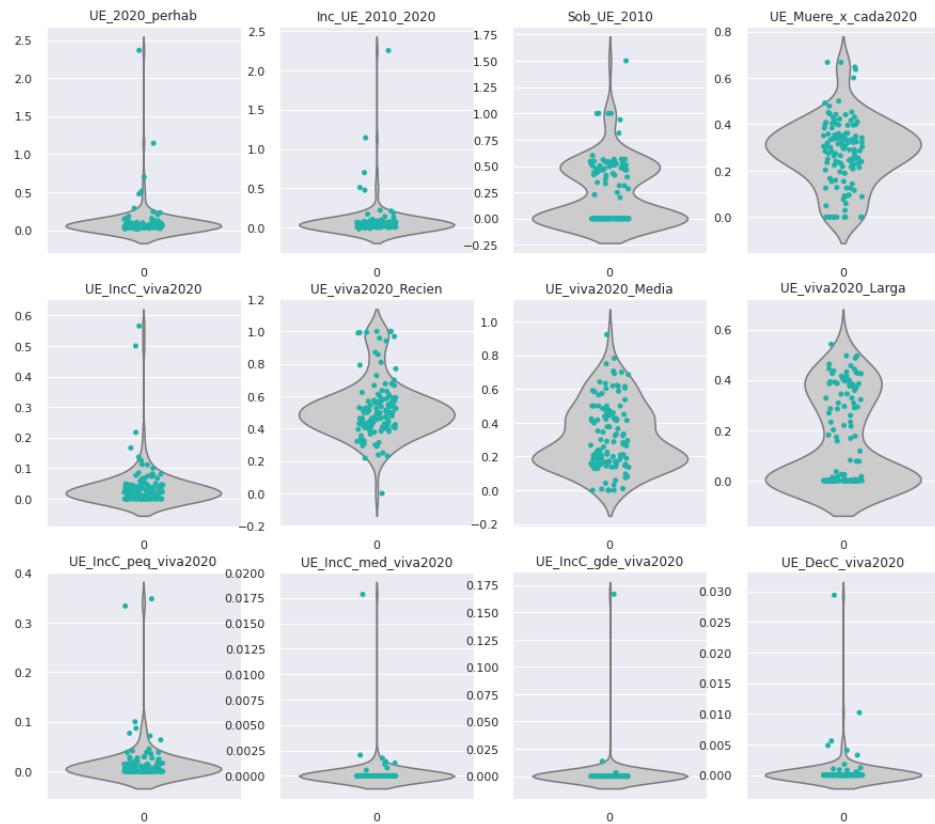


Figura 4.5: Dispersion variables originales. Base DENUE 126 localidades Nayarit.

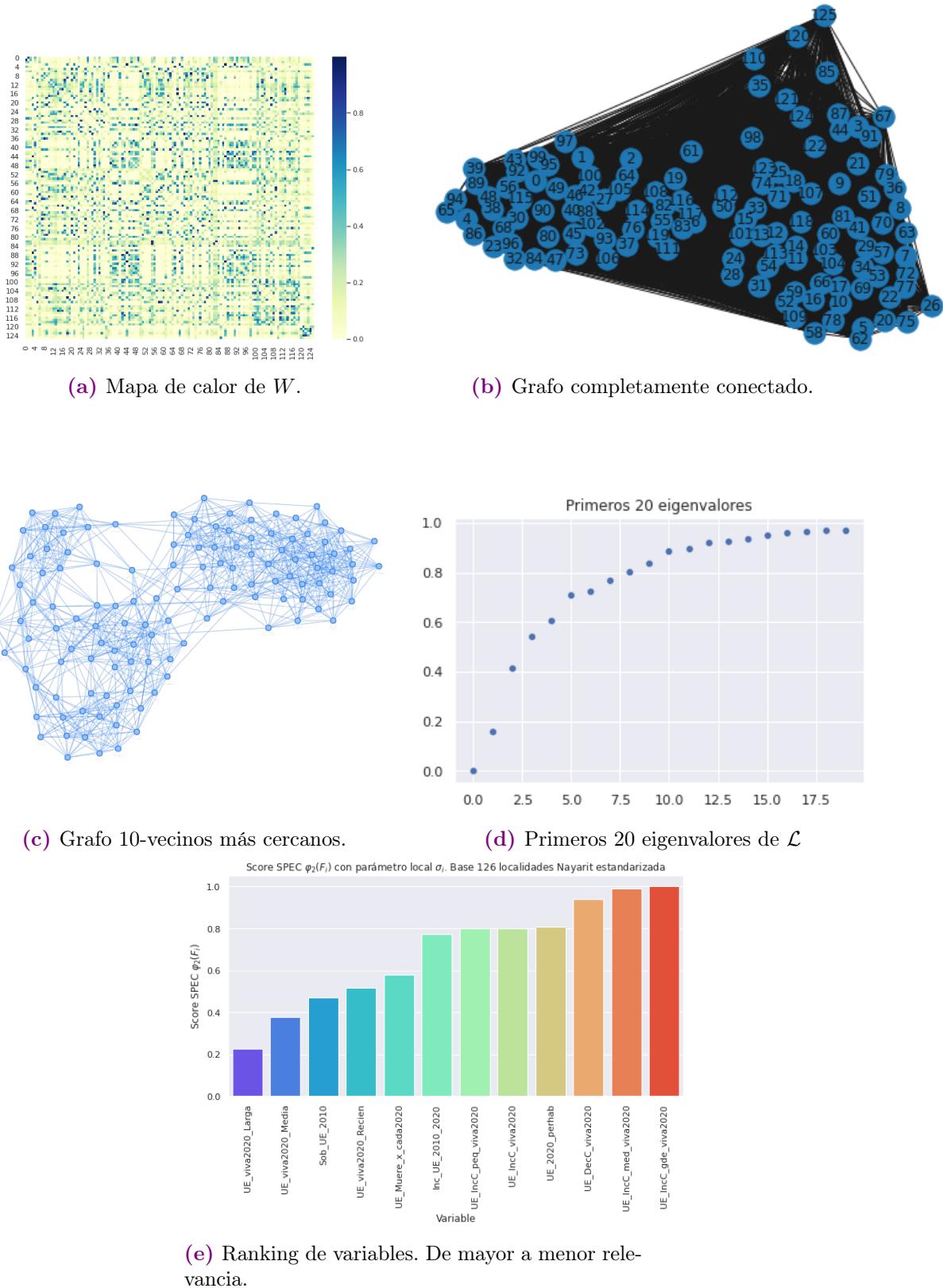


Figura 4.6: Resultados algoritmo SPEC con parámetro local σ_i . Base DENUE 126 localidades Nayarit estandarizada.

El siguiente conjunto de variables a las que se le aplicará el algoritmo SPEC son las variables socio-demográficas del CPV. Cabe señalar que de las 126 localidades a clusterizar, 4 no cuentan con información completa por lo tanto no se consideran en este grupo de variables.

MUN_2020	NOM_MUN_2020	LOC_2020	NOM_LOC_2020	POBTOT_2020	UE_Micro_Vivas2020
0	4	Compostela	605 El Crucero de Chacala	11	26
1	12	San Blas	24 Las Islitas	28	32
2	17	Tepic	427 El Crucero	10	7
3	20	Bahía de Banderas	236 Desarrollo Punta Mita	246	6

Figura 4.7: Localidades sin variables socio-demográficas.

Los resultados se pueden observar en la figura 4.9, en 4.9b se observa la representación del grafo completamente conectado, se puede apreciar que se trata de un conjunto de observaciones muy uniformes ya que no se distinguen grupos bien formados, incluso en la Figura 4.9c el grafo de los 10-vecinos más cercanos no se aprecia una clara diferencia de grupos.

Asimismo, los eigenvalores indican que posiblemente sólo hay un clúster dado que el primer eigenvalor es cero seguido de una larga brecha o gap respecto al siguiente eigenvalor.

Finalmente, en la figura 4.9e se puede notar que todas las variables influyen de manera similar en la clusterización, lo cual esta totalmente acorde a lo visto en el grafo y en la dispersión uniforme de las variables de la Figura 4.8, es decir, no hay una variable que discrimine fuertemente las instancias y por lo tanto no hay grupos claros.

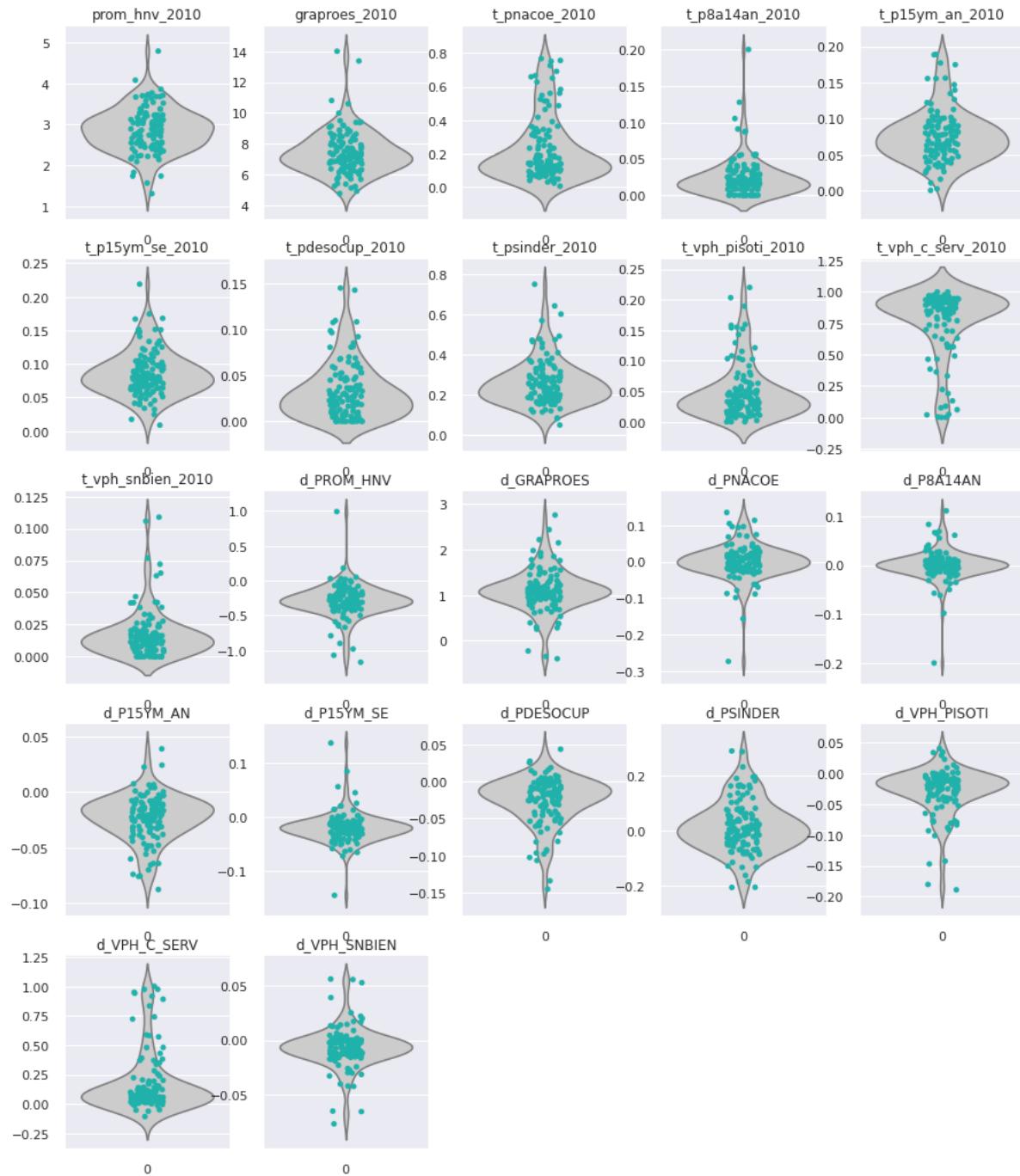


Figura 4.8: Dispersión variables originales. Base CPV 122 localidades Nayarit.

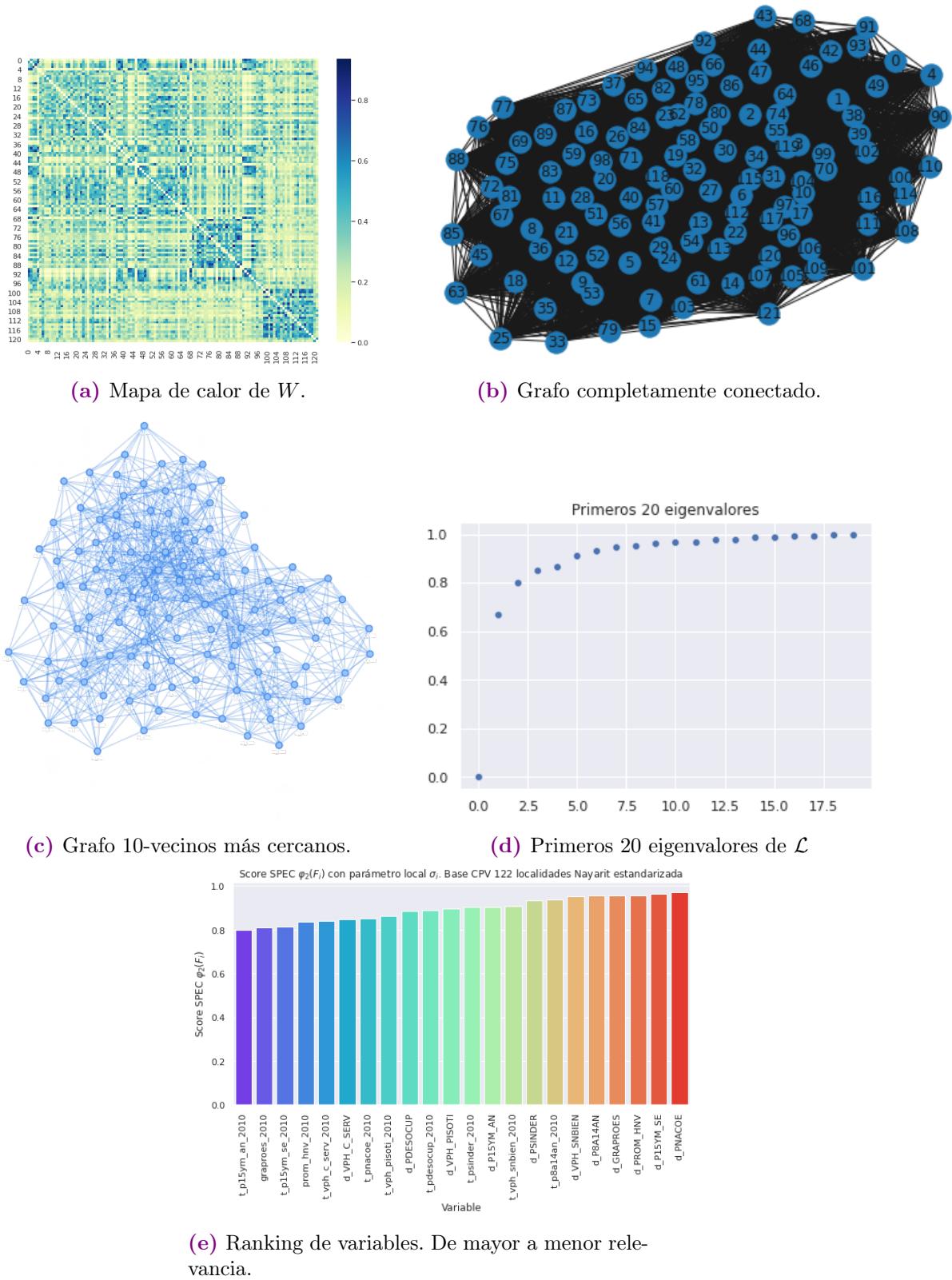


Figura 4.9: Resultados algoritmo SPEC con parámetro local σ_i . Base CPV 122 localidades Nayarit estandarizada.

Finalmente, se aplica el algoritmo SPEC a la base de datos de las 122 localidades considerando sus variables económicas y socio-demográficas.

Los resultados se pueden observar en la figura 4.10, se puede apreciar que al considerar ambos conjuntos de observaciones se obtiene un grafo muy uniforme, similar al grafo del conjunto de variables del CPV, lo cual indica que al combinar los dos conjuntos, las variables del CPV al ser mayor en cantidad tienen mayor influencia en la estructura del grafo y en consecuencia las variables del DENUE pierden relevancia.

En la figura 4.10e se puede notar que todas las variables, tanto del DENUE como del CPV, influyen de manera similar en la clusterización, lo cual es totalmente acorde a lo visto en el grafo, es decir, no hay una variable que discrimine fuertemente las instancias y por lo tanto no hay grupos claros.

Por lo anterior, se decide formar clústeres para cada conjunto de variables por separado, de manera que tendremos grupos que describan comportamientos similares considerando sólo las variables económicas del DENUE y grupos que describan comportamiento similares considerando sólo las variables socio-demográficas del CPV, lo cual beneficiará el proceso de interpretación de los resultados.

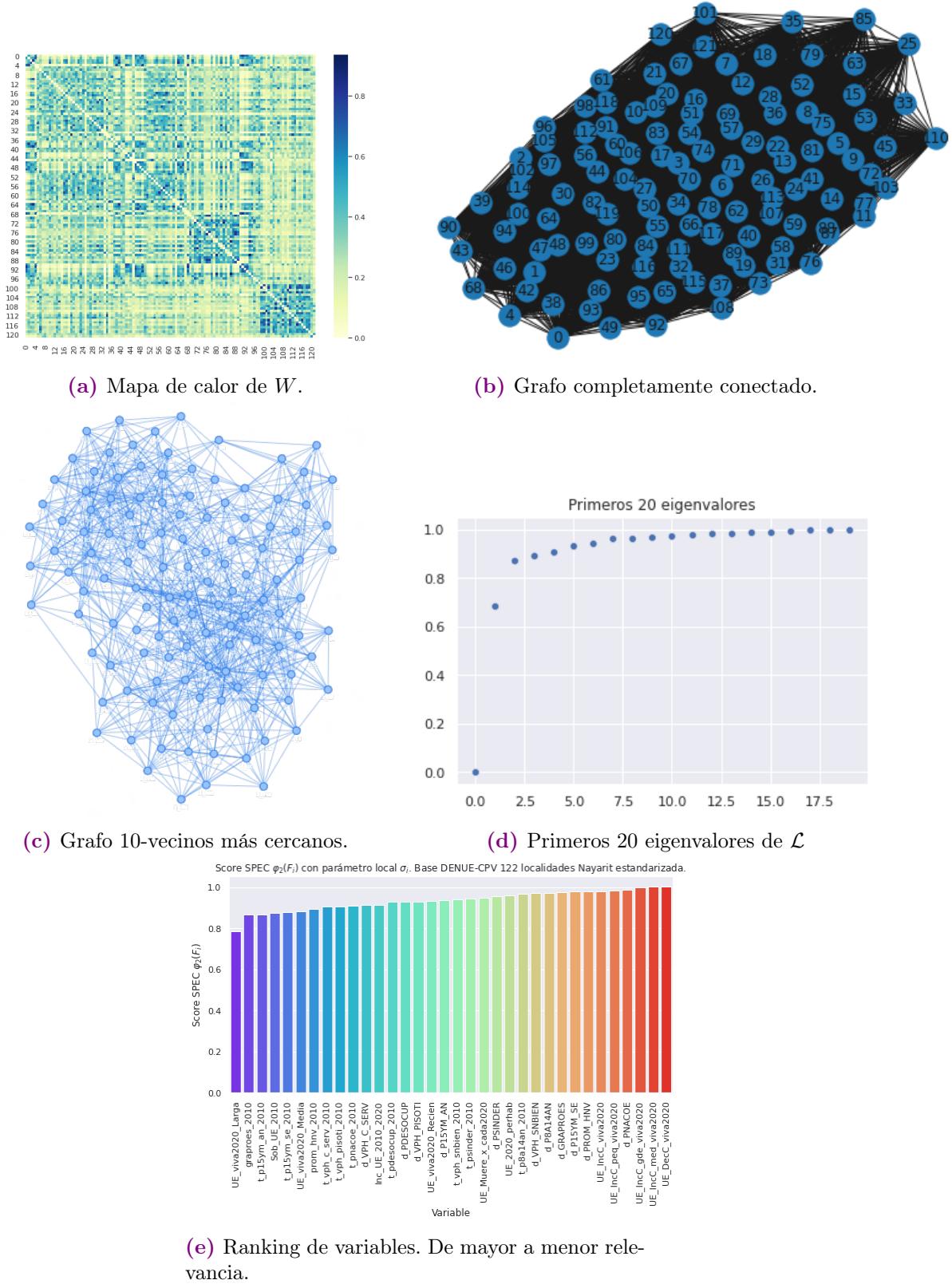


Figura 4.10: Resultados algoritmo SPEC con parámetro local σ_i . Base DENUE-CPV 122 localidades Nayarit estandarizada.

4.1.3. *Clustering* Espectral

Grafo completamente conectado, 12 variables DENUE

Se comienza aplicando el algoritmo de *clustering* espectral al grafo completamente conectado que contempla las 12 variables del DENUE. Se grafican las 5 primeras dimensiones del *embedding* espectral ya que el heurístico eigengap indica que los 5 gaps más grandes se encuentran en los eigenvalores [2, 1, 3, 5, 4], ordenados de mayor a menor, es decir, el gap en 2 es mayor al gap en 1 y así sucesivamente.

Posteriormente, se obtiene el gráfico de dispersión por pares del *embedding* espectral con la finalidad de visualizar posibles grupos por medio de las curvas de nivel y las estimaciones de densidad (Figura 4.11). Se pueden apreciar al menos 3 grupos, dos de ellos densamente poblados (ver dimensión 2) y el tercer grupo con menor densidad (ver dimensión 1). En la dimensión 5 no se logra apreciar un efecto discriminante sobre las observaciones, lo cual sugiere que esta dimensión no sea considerada, no obstante, se realiza el ejercicio con la finalidad de comparar resultados.

En la Figura 4.12 se muestra la agrupación resultante de aplicar *k-means* con $k = 5$ sobre el *embedding* espectral, se observa que el grupo amarillo es innecesario ya que sus elementos pueden pertenecer al clúster rosa o al clúster verde y su distinción no resalta alguna característica particular de sus elementos, lo anterior sugiere que se aplique *k-means* con $k = 4$ sobre las 4 primeras dimensiones del *embedding*.

En la agrupación con $k = 4$ se obtuvo un resultado interesante, se forman 3 grupos (verde, azul y rosa) cuya interpretación en las 5 variables más relevantes es clara y el cuarto grupo (rojo) agrupa las observaciones que presentan los valores más altos en las últimas 7 variables cuya relevancia es menor. En la dispersión de las ultimas 7 variables observamos que sólo un número reducido de instancias presentan valores altos y el resto presenta valores muy similares. Por lo tanto, el clúster rojo esta caracterizado por agrupar las observaciones con comportamiento atípico en las últimas 7 variables.

Asimismo, se realiza el agrupamiento con $k = 3$, la desventaja respecto a la agrupación con $k = 4$ es que se pierde el poder de distinguir las localidades en las que ha

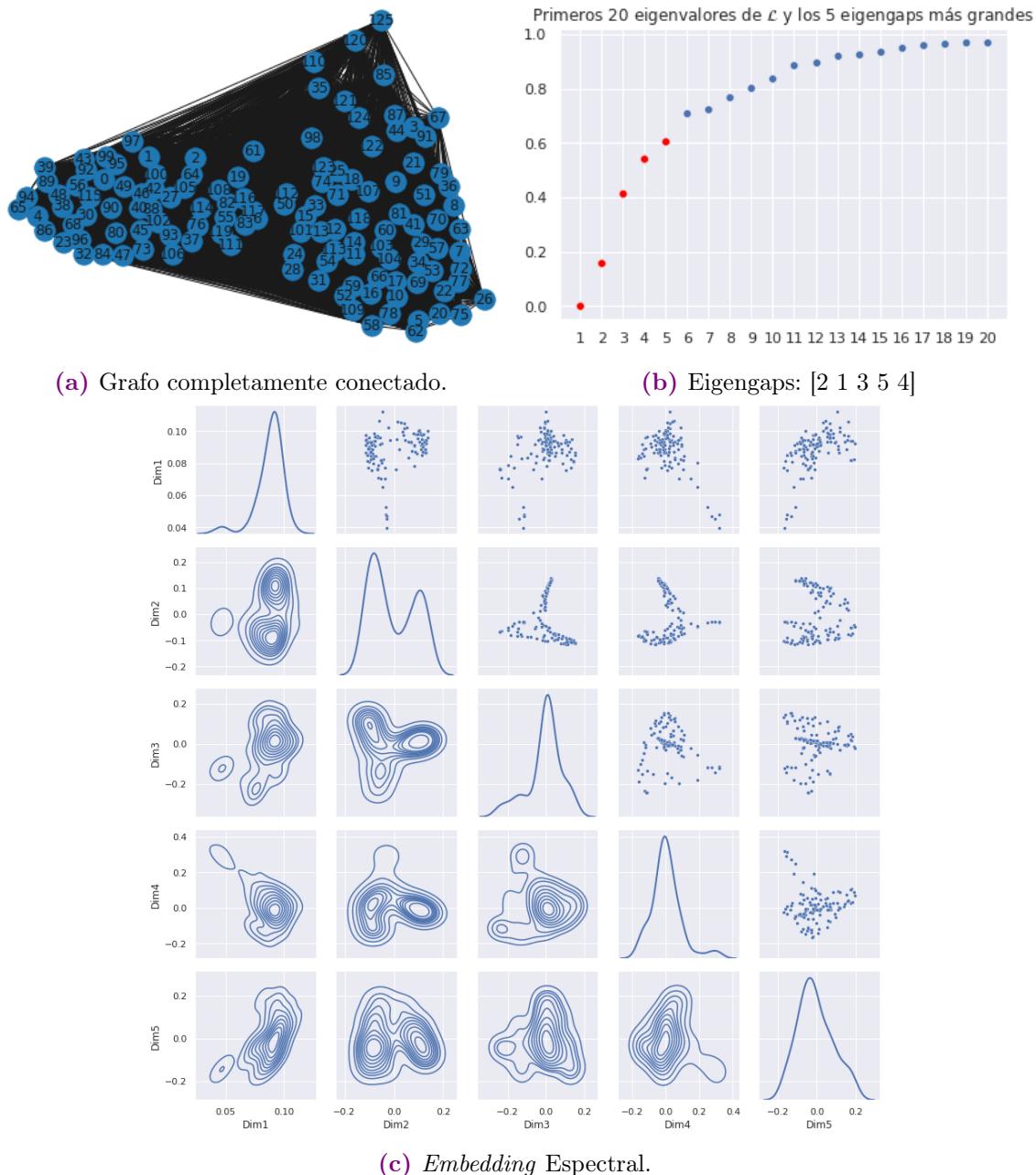


Figura 4.11: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base de 12 variables DENUE, 126 localidades Nayarit estandarizada.

habido una mayor aumento de UE de recién creación (creadas de 2019 en adelante) pero que no han presentado crecimientos en cuanto a número de personal, ya que en este caso se unen el clúster rojo con el clúster rosa del agrupamiento con $k = 4$.

Con base en lo anterior, bajo el escenario donde se consideran las 12 variables del DENUE, se elige la agrupación generada con $k = 4$.

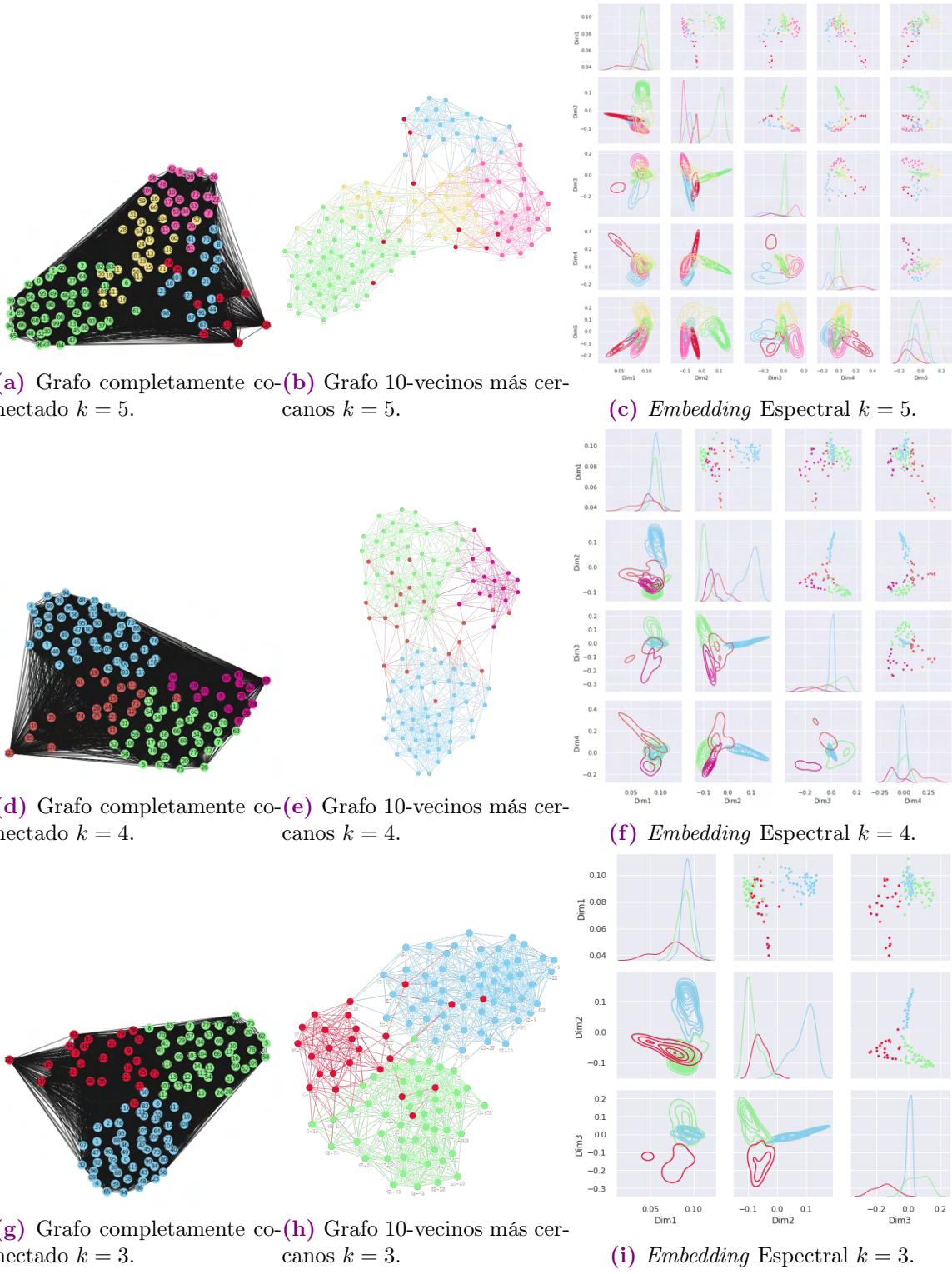


Figura 4.12: Resultados *clustering* espectral con parámetro local σ_i . Base de 12 variables y 126 localidades Nayarit estandarizada.

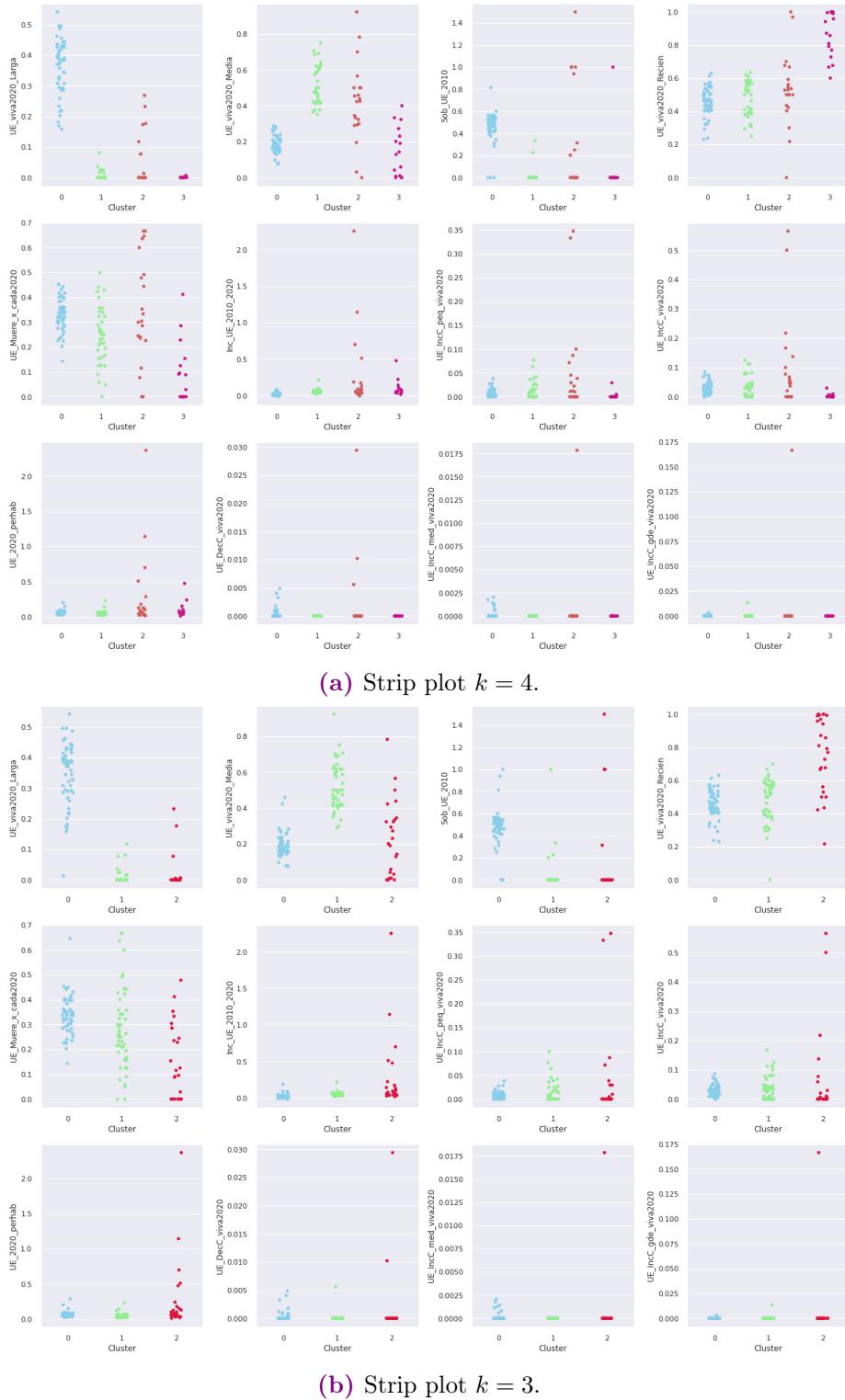


Figura 4.13: Distribución de clústeres. Base de 12 variables y 126 localidades Nayarit estandarizada.

En este punto la pregunta es qué sucedería si removemos las 7 variables de menor relevancia que están provocando la formación del clúster rojo. En la siguiente sección se presentan los resultados de realizar el ejercicio de eliminar de una en una las variables de menor relevancia y se podrá observar si mejoran los criterios de ajuste Silhouette, Calinski Harabasz y Davies Bouldin.

Grafo completamente conectado, eliminando variables de menor relevancia DENUE.

Para visualizar el efecto de remover las variables de menor relevancia se decide realizar un ciclo que en cada iteración remueva la variable de menor relevancia de acuerdo al ranking del algoritmo SPEC. Además, para cada iteración se obtendrá el heurístico eigengap y los criterios de ajuste Silhouette, Calinski Harabasz y Davies Bouldin; lo anterior con la finalidad de observar la mejora en cuanto a la composición de los clústeres.

Score	Dimensiones	V. Eliminada	Eigengap Top5	Número de Clusters														
				2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Silhouette	(126, 12)	Ninguna	[2 1 3 5 4]	0.78	0.64	0.57	0.48	0.43	0.45	0.41	0.39	0.40	0.39	0.38	0.36	0.38	0.38	
	(126, 11)	UE_IncC_gde_viva2020	[2 1 3 5 4]	0.75	0.64	0.56	0.48	0.45	0.45	0.41	0.40	0.39	0.37	0.37	0.35	0.38	0.37	
	(126, 10)	UE_IncC_med_viva2020	[2 1 3 5 4]	0.75	0.62	0.52	0.44	0.45	0.45	0.39	0.37	0.37	0.35	0.35	0.34	0.35	0.34	
	(126, 9)	UE_DecC_viva2020	[2 1 3 5 10]	0.75	0.62	0.53	0.45	0.46	0.45	0.41	0.39	0.38	0.39	0.37	0.36	0.37	0.37	
	(126, 8)	UE_2020_perhab	[2 3 1 4 6]	0.75	0.63	0.54	0.44	0.45	0.45	0.40	0.39	0.40	0.40	0.37	0.36	0.36	0.33	
	(126, 7)	UE_IncC_viva2020	[2 3 1 6 4]	0.75	0.62	0.52	0.48	0.48	0.46	0.40	0.41	0.41	0.38	0.38	0.38	0.39	0.37	
	(126, 6)	UE_IncC_paq_viva2020	[2 3 6 1 4]	0.75	0.58	0.59	0.51	0.52	0.44	0.41	0.43	0.42	0.43	0.42	0.44	0.42	0.39	
Calinski Harabasz	(126, 5)	Inc_UE_2010_2020	[3 6 4 2 1]	0.76	0.66	0.59	0.53	0.55	0.47	0.46	0.44	0.44	0.48	0.44	0.45	0.48	0.42	
	(126, 12)	Ninguna	[2 1 3 5 4]	791.24	266.77	132.76	91.34	70.16	64.48	51.02	41.75	38.16	32.85	28.15	24.87	25.59	23.05	
	(126, 11)	UE_IncC_gde_viva2020	[2 1 3 5 4]	804.71	263.32	135.49	89.75	72.64	64.51	50.32	41.98	37.31	30.99	26.99	23.56	25.37	22.64	
	(126, 10)	UE_IncC_med_viva2020	[2 1 3 5 4]	775.29	237.76	136.59	80.88	70.29	62.27	46.29	39.65	34.34	29.12	25.44	22.26	22.68	19.96	
	(126, 9)	UE_DecC_viva2020	[2 1 3 5 10]	796.67	241.91	150.09	80.64	70.44	63.02	48.00	41.47	34.97	33.44	27.44	23.34	24.51	22.07	
	(126, 8)	UE_2020_perhab	[2 3 1 4 6]	803.69	259.56	159.47	74.84	66.40	62.64	46.80	40.60	36.83	34.29	27.72	25.21	22.35	19.21	
	(126, 7)	UE_IncC_viva2020	[2 3 1 6 4]	792.55	233.61	142.10	91.41	74.20	65.62	46.66	43.45	37.77	30.49	27.26	26.11	23.48	20.80	
Davies Bouldin	(126, 6)	UE_IncC_paq_viva2020	[2 3 6 1 4]	798.66	203.57	182.26	106.12	93.88	62.94	47.51	44.60	40.59	39.30	32.99	32.18	27.24	23.29	
	(126, 5)	Inc_UE_2010_2020	[3 6 4 2 1]	893.45	295.00	201.38	129.69	109.67	67.23	57.75	47.91	44.02	47.59	36.24	33.31	35.46	26.59	
	(126, 12)	Ninguna	[2 1 3 5 4]	0.34	0.52	0.74	0.76	0.80	0.77	0.83	0.84	0.85	0.93	0.97	0.90	0.88	0.87	
	(126, 11)	UE_IncC_gde_viva2020	[2 1 3 5 4]	0.34	0.54	0.71	0.74	0.78	0.77	0.83	0.83	0.88	0.93	0.97	0.98	0.88	0.90	
	(126, 10)	UE_IncC_med_viva2020	[2 1 3 5 4]	0.34	0.56	0.67	0.80	0.79	0.77	0.86	0.88	0.89	0.92	1.01	0.97	0.94	0.99	
	(126, 9)	UE_DecC_viva2020	[2 1 3 5 10]	0.33	0.53	0.65	0.82	0.79	0.77	0.85	0.89	0.90	0.88	0.92	0.98	0.90	0.95	
	(126, 8)	UE_2020_perhab	[2 3 1 4 6]	0.33	0.51	0.64	0.84	0.84	0.78	0.87	0.90	0.86	0.84	0.91	0.94	0.99	0.96	
	(126, 7)	UE_IncC_viva2020	[2 3 1 6 4]	0.33	0.54	0.68	0.79	0.79	0.75	0.87	0.82	0.89	0.93	1.00	0.96	0.97	0.94	
	(126, 6)	UE_IncC_paq_viva2020	[2 3 6 1 4]	0.33	0.64	0.57	0.71	0.70	0.79	0.84	0.81	0.81	0.78	0.82	0.78	0.83	0.92	
	(126, 5)	Inc_UE_2010_2020	[3 6 4 2 1]	0.31	0.43	0.55	0.66	0.67	0.80	0.77	0.82	0.80	0.74	0.80	0.84	0.73	0.86	

Figura 4.14: Resultados al eliminar las variables de menor relevancia Base DENUE 126 localidades Nayarit estandarizada.

Los resultados se presentan en la figura 4.14, se observa que el criterio Silhouette y Calinski mejoran muy ligeramente al eliminar las 7 variables de menor relevancia y el criterio Davies Bouldin es el que presenta una mayor mejora, sin embargo, los tres criterios apuntan hacia elegir $k = 2$ clústeres. No obstante esto provocaría una mayor

generalización de los grupos y se perderían de vista ciertas características interesantes, por lo tanto la opción de $k = 2$ clústeres se descarta.

Por otro lado, el heurístico eigengap indica que al remover las variables menor relevancia, el número de clústeres óptimos se encontrará en $k = [3, 6, 4, 2, 1]$. En la Figura 4.15 se muestra la dispersión del *embedding* espectral de las primeras 6 dimensiones, en este se logran visualizar hasta 3 o incluso 4 posibles grupos. A continuación se muestra el resultado de aplicar el algoritmo de *clustering* con $k = 6$, $k = 4$ y $k = 3$.

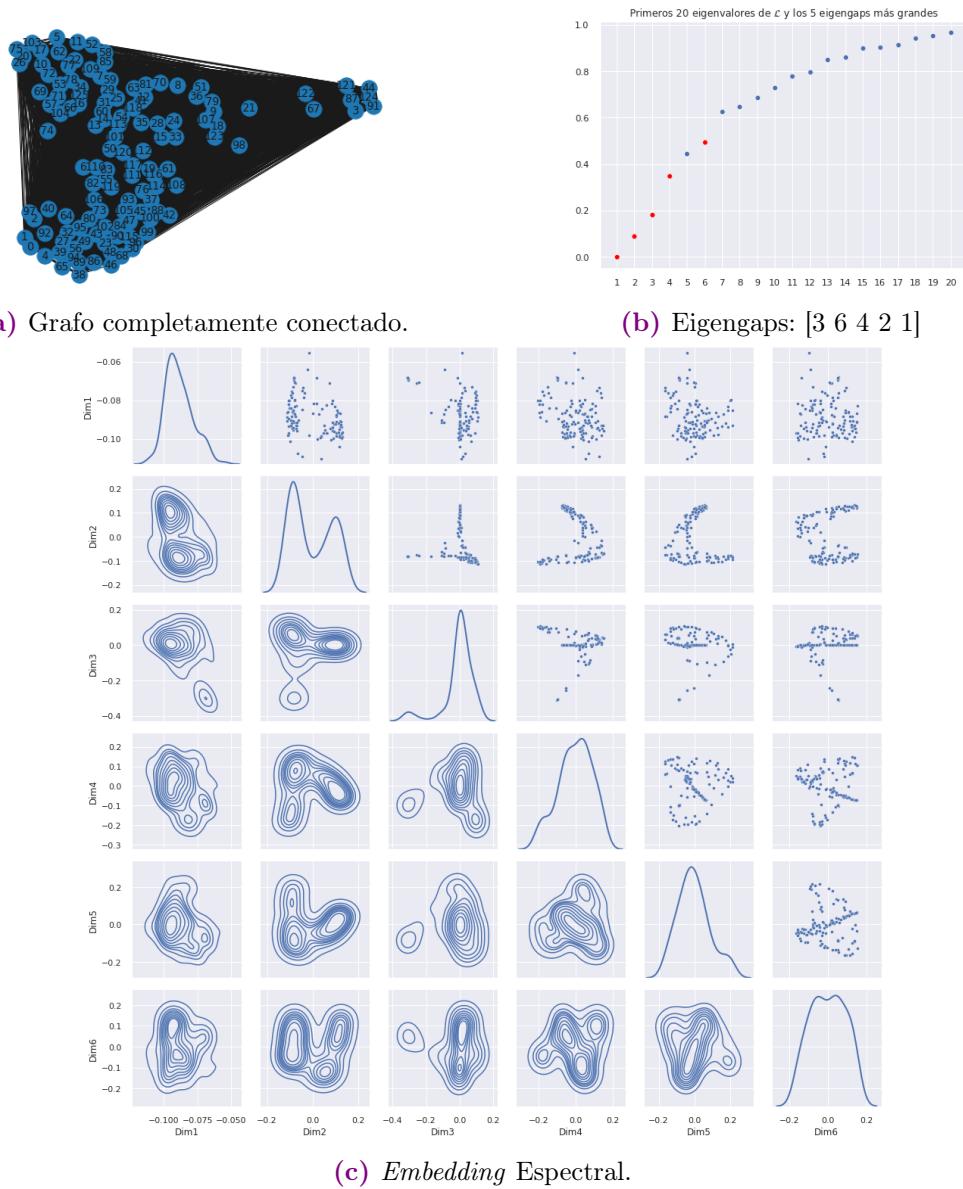


Figura 4.15: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base de 5 variables DENU, 126 localidades Nayarit estandarizada.

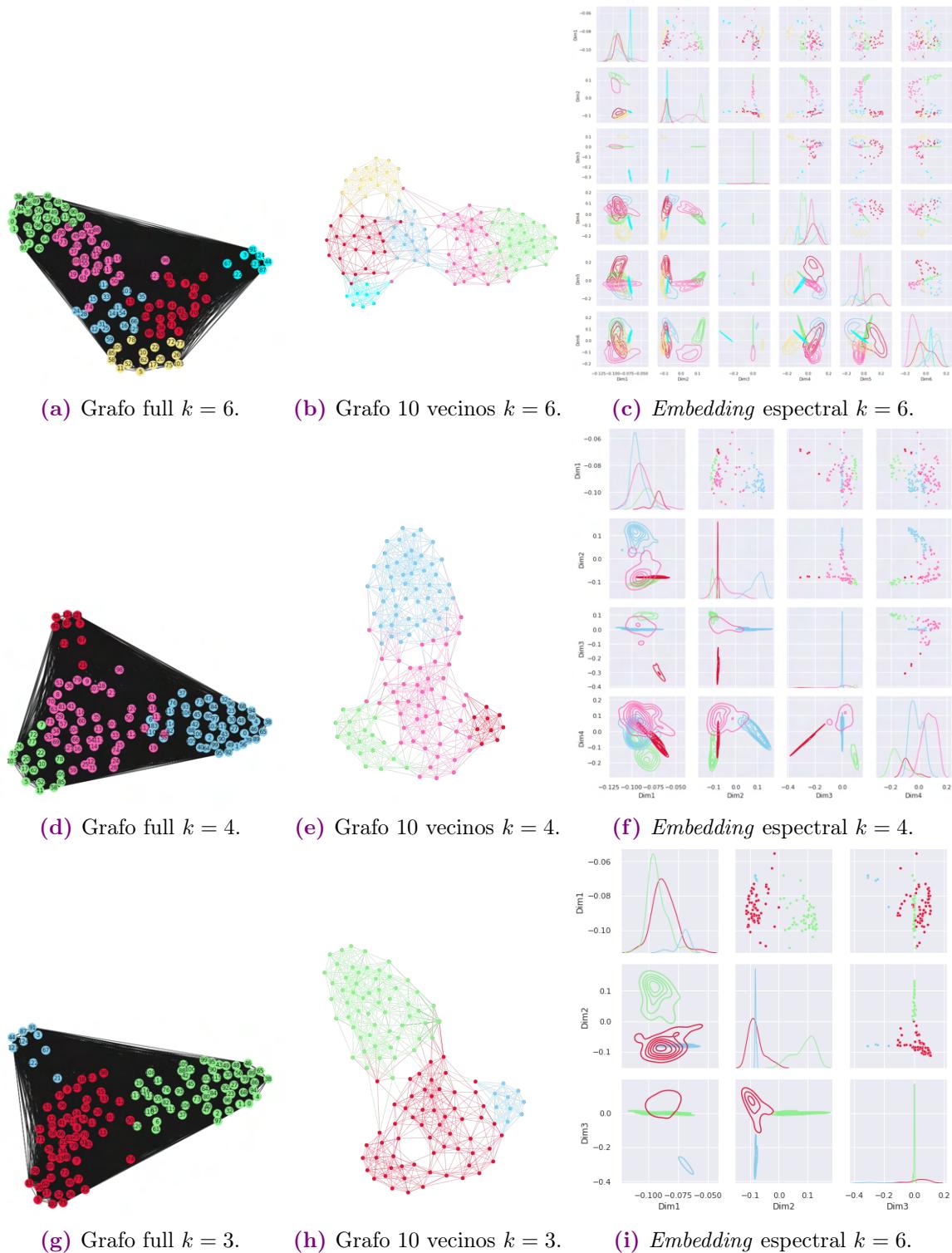


Figura 4.16: Resultados *clustering* espectral con parámetro local σ_i . Base de 5 variables y 126 localidades Nayarit estandarizada.

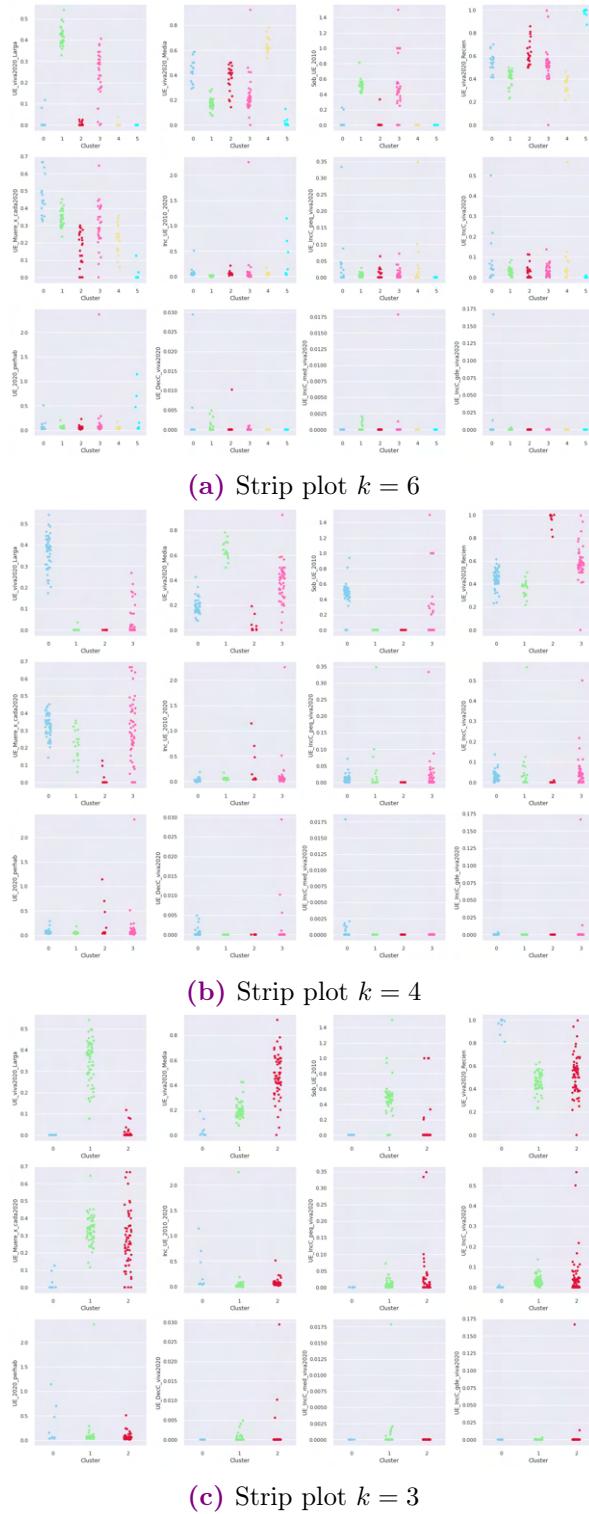


Figura 4.17: Distribución de clústeres. Base de 5 variables y 122 localidades Nayarit estandarizada..

En los clústeres generados con las variables más relevantes se observa mayor claridad en cuanto a los grupos de localidades por antigüedad, sin embargo, las últimas 7 variables, que son las que miden los incrementos de personal en los micro negocios, no revelan algún grupo con un comportamiento que destaque del resto, es decir, todos los grupos se distribuyen de manera similar en las 7 variables. Lo anterior hace que la interpretación sea más complicada, razón por la cual se decide elegir como agrupación final el *clustering* generado a partir de las 12 variables con $k = 4$.

Eigensearch o Búsqueda iterativa del Eigengap. Clústeres DENUE.

Después de haber elegido como versión final la agrupación con $k = 4$ y 12 variables del DENUE, aplicamos el algoritmo Eigensearch sobre cada uno de los clústeres generados, con la finalidad de verificar si dentro de cada clúster hay posibilidad de encontrar un nuevo clúster. Si el gap más grande de los eigenvalores de la matriz Laplaciana de cada clúster ocurre en el primer eigenvalor entonces no hay más clústeres por modelar.

En la gráficas de la Figura 4.18 se puede observar que el gap en el primer eigenvalor es considerablemente mayor al resto de gaps, lo cual indica que ya no hay más clústeres por modelar.

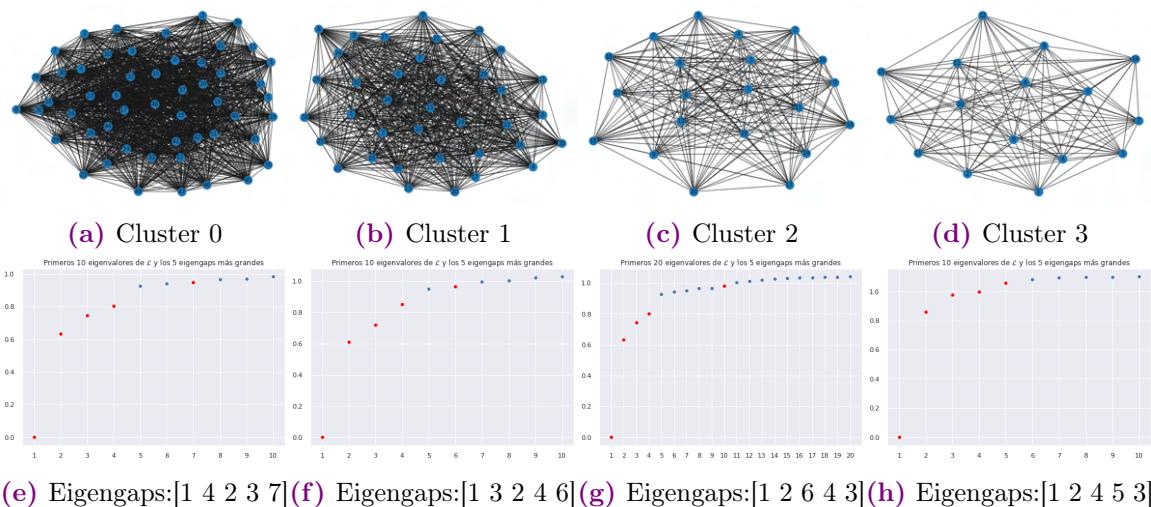


Figura 4.18: Eigensearch.

Interpretación Clústeres DENUE

Los 4 clústeres generados con las variables del DENUE para el estado de Nayarit quedan distribuidos de la siguiente manera:

Cluster	Localidades	%
Azul	52	41 %
Verde	37	29 %
Rojo	21	17 %
Rosa	16	13 %
Total	126	100 %

Tabla 4.2: Clústeres DENUE Nayarit.

A continuación se realiza una interpretación de los mismos con base en la dispersión de las variables originales y sus cuatiles poblacionales, además, se muestra cómo a partir del *embedding* espectral se logra resumir esta información por medio de la distribución de los puntos en el espacio de dimensión reducida.

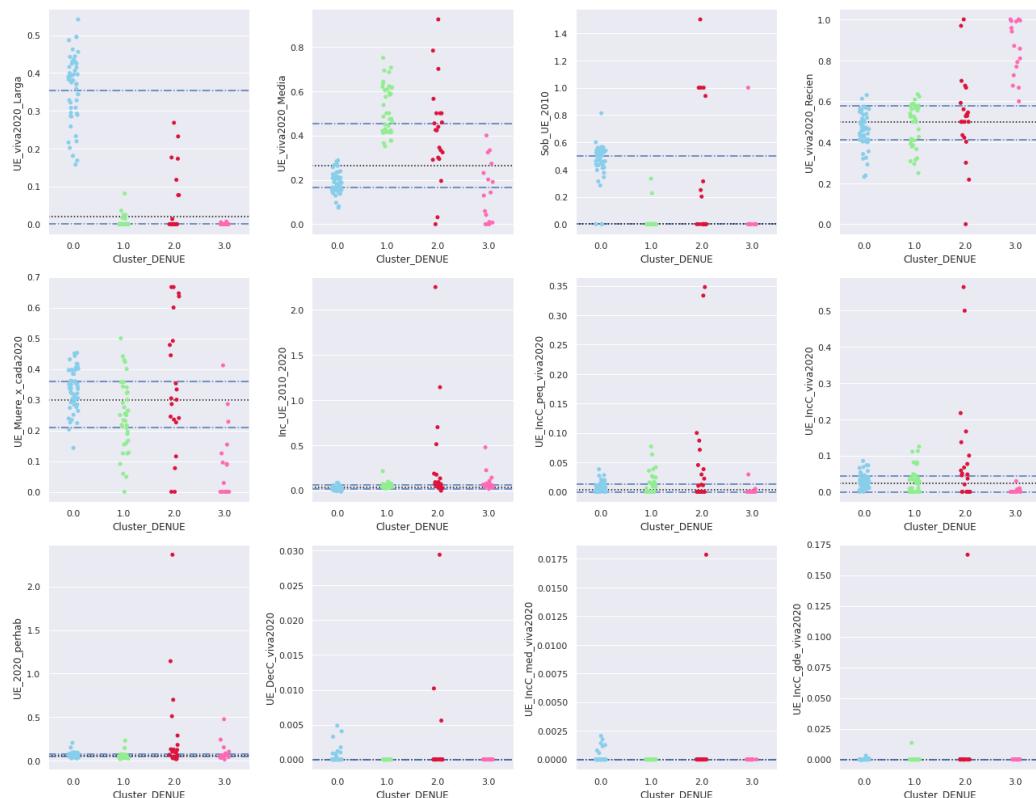


Figura 4.19: Dispersion clústeres DENUE y cuantiles .25, .50 y .75 poblacionales.

En la Figura 4.20 se muestra la correlación lineal entre las variables y las dimensiones del *embedding*.

La dimensión 1 presenta una débil correlación con todas las variables pero en general valores más negativos implican mayor crecimiento del sector micro. Recorremos que la dimensión 1 corresponde al eigenvector del primer eigenvalor, cuyo valor siempre es 0, por lo tanto se espera que sea poco informativa.

La dimensión 2 esta fuertemente relacionada positivamente a localidades con antigüedades largas cuya tasa de sobre-vivencia es alta y negativamente con antigüedades medias o bajas.

La dimensión 3 distingue a las localidades de antigüedad media y baja, valores negativos se relacionan con las localidades más jóvenes.

La dimensión 4, se relaciona positivamente a localidades que han presentado crecimientos en tamaño de personal, pasaron de ser micro a pequeñas empresas.

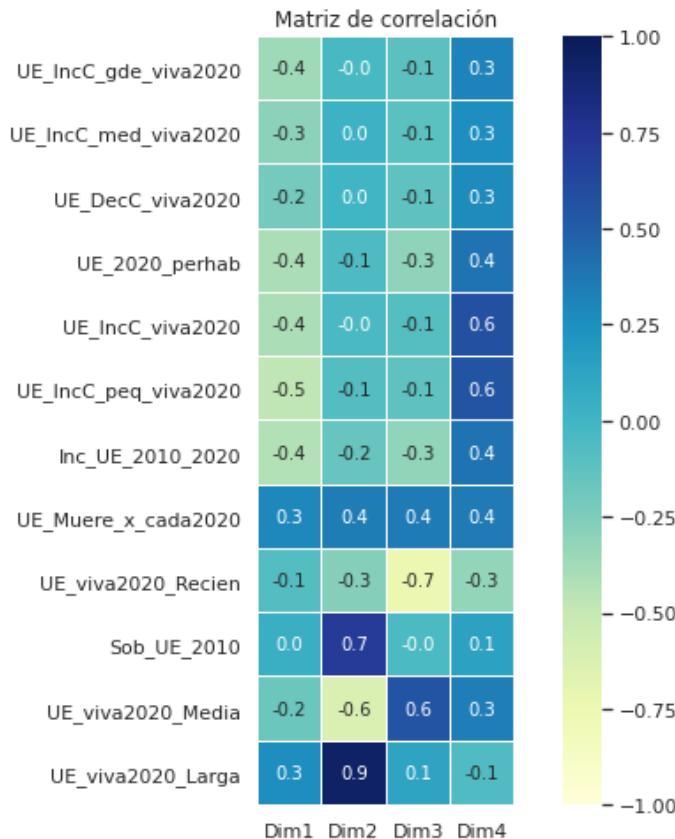


Figura 4.20: Matriz de correlación variables originales vs. *embedding*.

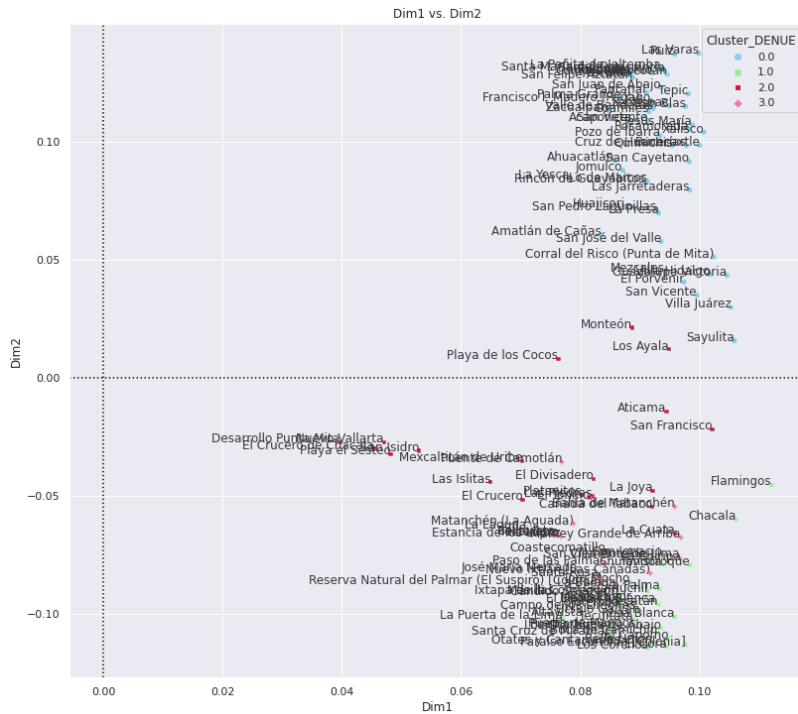


Figura 4.21: Dim1 vs. Dim2

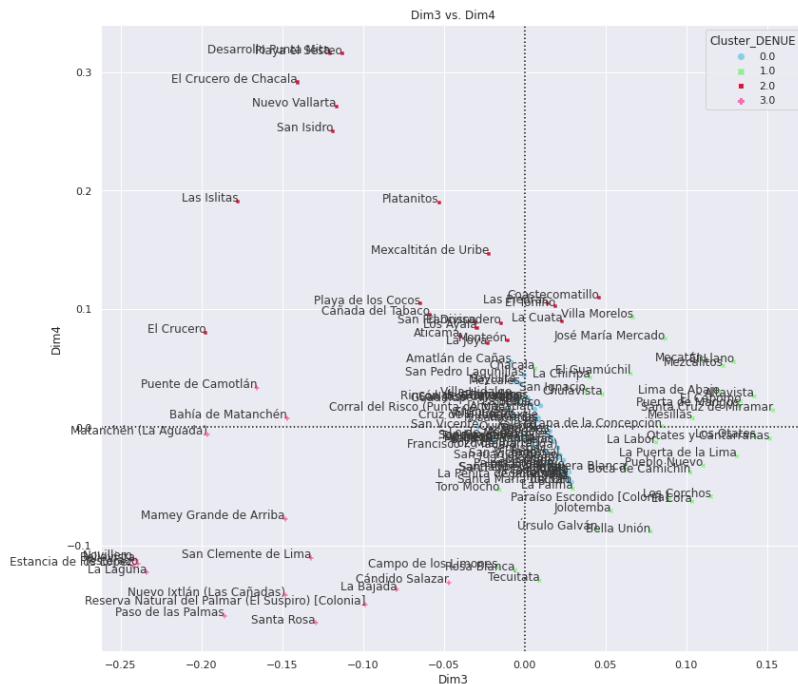


Figura 4.22: Dim3 vs. Dim4

Con base en los gráficos mostrados, se presenta a continuación la interpretación por clúster:

Clúster Azul:

- Se caracteriza por incluir localidades que presentan las mayores antigüedades. En la mitad de las localidades de este clúster el porcentaje de UE dadas de alta antes de 2014 y que sobreviven en 2020 es de al menos 38 %.
- La tasa de sobre-vivencia a largo plazo es alta (45 % aproximadamente), sobre-sale del resto de localidades.
- La proporción de UE de recién creación (dadas de alta a partir de 2019) se encuentra en los niveles normales en comparación al resto (50 % aprox.), por lo que también presenta un escenario favorecedor para nuevos negocios.
- El incremento en el número de UE por habitante es de los más bajos, es decir, el crecimiento de UE se esta dando a la par del crecimiento en la población.
- En este clúster se han dado algunos crecimientos de empresas micro a pequeñas y medianas. Las localidades con mayores incrementos de este tipo son: Cruz de Huanacaxtle, Rincón de Guayabitos, Las Jarretaderas, Sayulita, Mezcales, Bucerías, Jalcocotán, Lo de Marcos y San Blas.

Todo lo anterior sugiere que las localidades que lo conforman han alcanzado un nivel de crecimiento y se ha mantenido con un nivel constante, entran y salen empresas pero no hay un crecimiento fuera de los niveles normales dado que la población también ha incrementado a la par.

Cluster Verde:

- Este clúster presenta características muy similares al clúster azul con la diferencia de que es más joven, en la mitad de las localidades de este clúster el porcentaje de UE dadas de alta de 2014 a 2018 y que sobreviven en 2020 es de al menos 50 %.
- En las localidad Flamingos y Chacala el incremento en el número de unidades por habitante es alto.

- Las localidades que han presentado crecimientos de empresas micro a pequeñas son: El Guamúchil, Tondoroque, La Chiripa, El Capomo, Lima de Abajo, Chacala, Higuera Blanca y Flamingos.

Cluster Rojo:

- Este clúster es el más disperso, hay de todo tipo de antigüedades, sin embargo, predominan las UE que se crearon de 2014 a 2018 y que sobreviven en 2020 (42 % aprox.)
- Presenta los incrementos en el número de UE por habitante más altos, en las localidades El Crucero de Chacala, Las islitas, Platanitos, Playa El Sesteo y Los Ayala. Esto significa que han aumentado el número de negocios micro a un ritmo mayor que el crecimiento de la población.
- En este clúster se han dado mayores crecimientos de empresas micro a pequeñas empresas. La localidades con mayores incrementos de este tipo son: Playa el Sesteo, Desarrollo Punta de Mita, Platanitos, La Cuata, Nuevo Vallarta, Coastecamatillo, El crucero de Chacala y San Franciso.

Cluster Rosa:

- Este clúster es el más joven y por tanto el de menor solidez, en todas sus localidades más del 60 % de las UE se crearon de 2019 en adelante.
- Algunas de las localidades presentan incrementos en el número de UE por habitante, Matanchén, La Laguna, Nuevo Ixtlán, Las Cañadas, Puente de Camotlán y La Bajada. Sin embargo, solo en una localidad se han registrado incrementos en el tamaño de personal (La Bajada), el resto no presenta incrementos.

Tomando en cuenta todo lo anterior, se puede concluir que las localidades que presentan un mayor potencial de crecimiento se encuentran en el clúster rojo, ya que éste ha registrado los mayores crecimientos de micro a pequeña empresa, sus incrementos de UE por habitante son superiores al resto de las localidades y cuenta

con cierto nivel de solidez al conformarse por algunas localidades de antigüedad larga y media. Después del clúster rojo, los siguientes con mayor potencial de crecimiento son el clúster azul y verde, ambos con situaciones similares pero diferente antigüedad. La última posición la ocupa el clúster rosa, considerando que es un grupo que aún le falta madurez dada su corta antigüedad y el poco crecimiento que ha registrado.

Con base en lo anterior, se propone el siguiente *ranking* por nivel de potencial del sector micro, donde 4 estrellas es el máximo nivel y una estrella en nivel mínimo:

Cluster DENUE	Potencial sector Micro	Categoría
Rojo	★★★	A
Azul	★★★	B
Verde	★★★	B
Rosa	★★	C

Tabla 4.3: *Ranking* Clústeres DENUE Nayarit.

Grafo completamente conectado, 22 variables CPV.

Para las variables del CPV se aplica la misma metodología que se aplicó en el *clustering* basado en las variables del DENUE. Contemplando un grafo completamente conectado que considera las 22 variables del CPV se grafican las 6 primeras dimensiones del *embedding* espectral, ya que el heurístico eigengap indica que los 5 gaps más grandes se encuentran en los eigenvalores [1,2,3,5,6], ordenados de mayor a menor.

En la figura 4.23, el gráfico de dispersión del *embedding* espectral muestra que a lo más existen dos grupos (ver dimensión 4), sin embargo, la mayoría de observaciones se concentran en un sólo grupo. Esto también se ve reflejado en el gráfico de los eigengaps, el primer gap es por mucho superior al resto, lo cual indica que sólo debe haber un grupo. Además, se puede observar que la dimensión 5 y 6 no tienen un efecto discriminante en las instancias, por lo cual, se decide sólo visualizar los clústeres que contemplan las primeras 2, 3 y 4 dimensiones del *embedding*.

En los 3 clústeres que se muestran en la Figura 4.24 se puede apreciar que las agrupaciones obtenidas son de difícil interpretación, ya que visualmente no se observan ningún grupo de instancias bien separadas del resto. En los 3 casos los grupos son muy semejantes y existe una superposición de los clústeres. Es importante resaltar que en los gráficos de violín de la Figura 4.25 se observa que al menos las últimas 8 variables los clústeres presentan distribuciones muy similares, su medianas no difieren, por lo que estas variables no proveen información útil para generar los grupos.

Por lo tanto, bajo el esquema de considerar las 22 variables no se obtuvieron agrupaciones relevantes, por lo tanto, se opta por verificar si al eliminar las variables de menor relevancia es posible encontrar grupos interesantes.

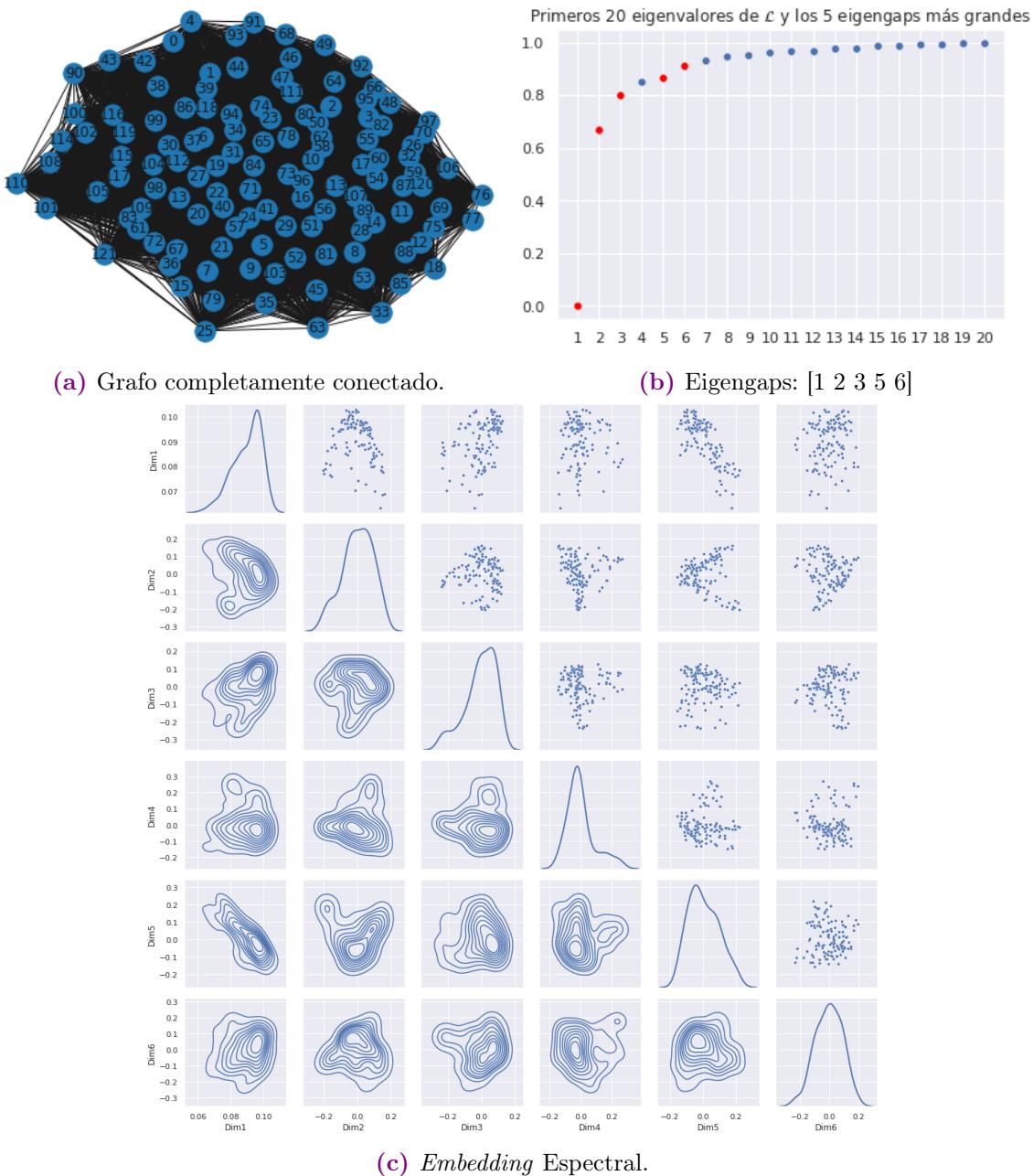


Figura 4.23: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base de 22 variables CPV, 122 localidades Nayarit estandarizada.

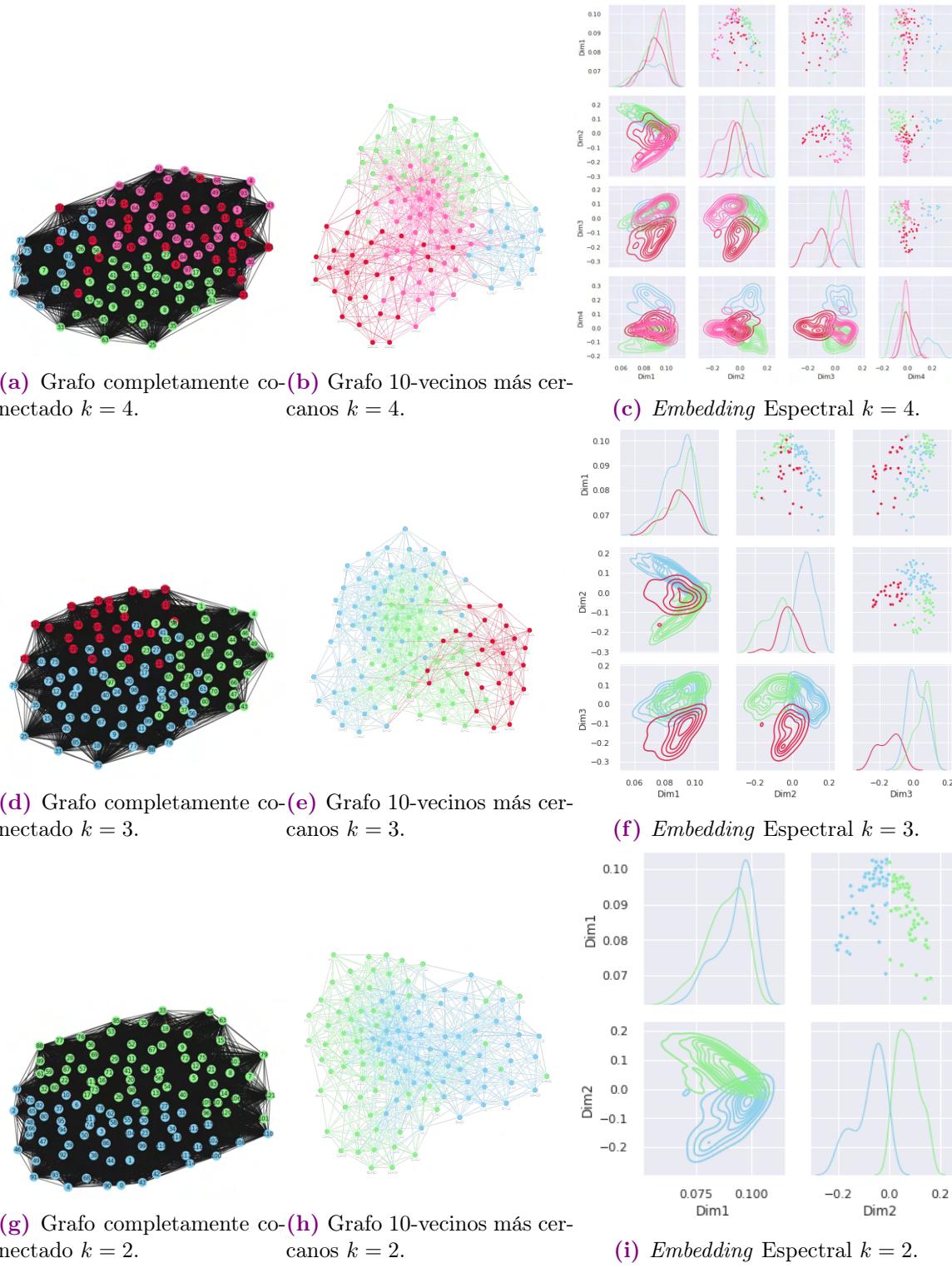
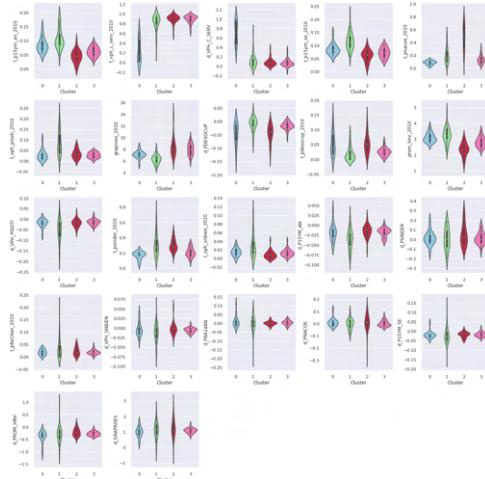
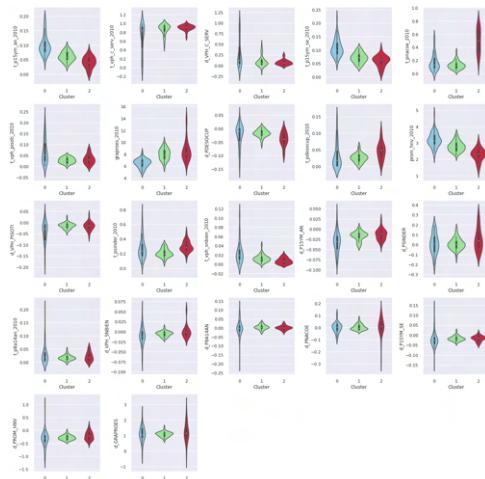


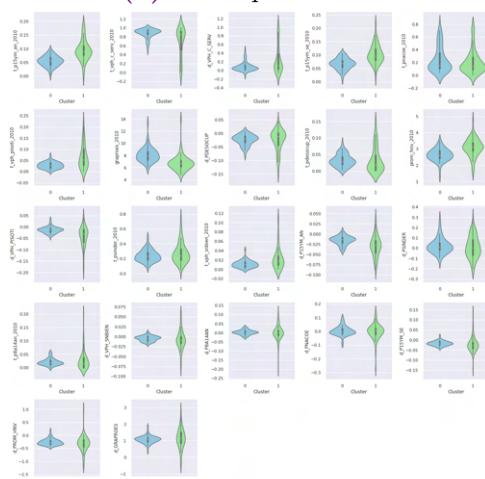
Figura 4.24: Resultados *clustering* espectral con parámetro local σ_i . Base de 22 variables del CPV y 122 localidades Nayarit estandarizada.



(a) Violin plot $k = 5$.



(b) Violin plot $k = 4$.



(c) Violin plot $k = 3$.

Figura 4.25: Distribución de clústeres por variable según *ranking* de relevancia. Base de 22 variables del CPV y 122 localidades Nayarit estandarizada.

Grafo completamente conectado, eliminando variables de menor relevancia CPV.

Se aplica el algoritmo de *clustering* removiendo de una en una las variables de menor relevancia de acuerdo al *ranking* SPEC. En la figura 4.26 se observa una mejora en los criterios de ajuste cuando se eliminan las 15 variables menos relevantes y cuando el número de clústeres es igual a 2,3,4 o 5. A continuación se muestran los resultados del *clustering*.

Valores	Dim	V. Eliminada	Eigengap Top5	Número de Clusters														
				2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Silhouette	(122, 22)	Ninguna	[1 2 3 5 6]	0.57	0.45	0.41	0.40	0.38	0.30	0.27	0.24	0.22	0.21	0.22	0.22	0.22	0.20	
	(122, 21)	d_GRAPROES	[1 2 3 5 6]	0.57	0.47	0.42	0.41	0.39	0.32	0.30	0.25	0.23	0.21	0.22	0.21	0.19	0.18	
	(122, 20)	d_PROM_HNV	[1 2 3 5 6]	0.57	0.48	0.43	0.40	0.38	0.32	0.30	0.26	0.24	0.22	0.21	0.20	0.21	0.20	
	(122, 19)	d_P15YM_SE	[1 2 3 5 6]	0.57	0.48	0.44	0.41	0.39	0.33	0.32	0.27	0.24	0.22	0.23	0.22	0.22	0.21	
	(122, 18)	d_PNACOE	[1 2 3 5 6]	0.60	0.49	0.43	0.40	0.39	0.32	0.31	0.27	0.24	0.22	0.23	0.22	0.22	0.20	
	(122, 17)	d_PBA14AN	[1 2 3 5 6]	0.60	0.49	0.44	0.41	0.40	0.32	0.30	0.27	0.25	0.25	0.24	0.22	0.21	0.22	
	(122, 16)	d_VPH_SNBIEN	[1 2 3 4 5]	0.61	0.48	0.45	0.42	0.39	0.32	0.31	0.29	0.27	0.27	0.26	0.25	0.22	0.24	
	(122, 15)	t_p08a14an_2010	[1 2 4 3 5]	0.62	0.47	0.45	0.41	0.40	0.33	0.31	0.29	0.28	0.28	0.27	0.24	0.25	0.26	
	(122, 14)	d_PSINDER	[1 2 4 6 3]	0.62	0.46	0.44	0.41	0.40	0.36	0.34	0.29	0.29	0.27	0.26	0.26	0.25	0.27	
	(122, 13)	d_P15YM_AN	[1 2 4 6 3]	0.63	0.44	0.45	0.40	0.39	0.35	0.33	0.32	0.32	0.30	0.28	0.27	0.28	0.28	
	(122, 12)	t_VPH_snbién_2010	[1 2 4 6 3]	0.62	0.43	0.44	0.41	0.40	0.35	0.34	0.33	0.32	0.30	0.29	0.29	0.29	0.27	
	(122, 11)	t_pander_2010	[1 3 2 5 6]	0.62	0.48	0.45	0.44	0.40	0.37	0.35	0.34	0.32	0.31	0.31	0.31	0.30	0.29	
	(122, 10)	d_VPH_PISOTI	[1 3 4 6 2]	0.63	0.42	0.46	0.42	0.43	0.40	0.39	0.36	0.35	0.33	0.32	0.31	0.29	0.30	
	(122, 9)	prom_hnv_2010	[1 4 3 6 5]	0.59	0.41	0.47	0.44	0.43	0.37	0.37	0.36	0.34	0.31	0.32	0.31	0.30		
	(122, 8)	t_pdesocup_2010	[1 4 5 2 3]	0.60	0.47	0.48	0.45	0.42	0.42	0.40	0.37	0.37	0.35	0.33	0.32	0.30		
Calinski Harabasz	(122, 7)	d_PDESOCUP	[1 5 4 2 3]	0.81	0.57	0.50	0.49	0.45	0.41	0.43	0.41	0.41	0.36	0.36	0.33	0.31	0.31	
	(122, 6)	draproces_2010	[1 4 5 2 9]	0.79	0.65	0.59	0.49	0.47	0.44	0.45	0.44	0.41	0.39	0.37	0.35	0.33	0.31	
	(122, 22)	Ninguna	[1 2 3 5 6]	246.24	125.48	91.90	71.38	53.96	36.17	25.90	20.47	16.85	14.53	13.39	12.21	11.32	9.69	
	(122, 21)	d_GRAPROES	[1 2 3 5 6]	233.32	134.68	94.01	74.11	56.61	38.69	29.57	21.70	17.22	14.82	13.63	12.18	10.90	9.27	
	(122, 20)	d_PROM_HNV	[1 2 3 5 6]	237.02	140.47	97.62	73.05	56.49	38.87	29.37	22.29	17.84	15.24	13.30	12.27	11.12	9.98	
	(122, 19)	d_P15YM_SE	[1 2 3 5 6]	232.14	137.42	96.31	73.74	57.71	40.30	30.83	23.01	17.83	15.57	14.31	12.95	11.85	10.48	
	(122, 18)	d_PNACOE	[1 2 3 5 6]	268.36	145.24	94.34	72.37	57.48	38.86	29.80	23.34	18.99	15.66	14.92	13.03	11.72	10.12	
	(122, 17)	d_PBA14AN	[1 2 3 5 6]	266.65	144.20	95.97	74.88	59.21	39.22	30.43	24.44	19.59	17.42	15.00	12.69	11.83	11.58	
	(122, 16)	d_VPH_SNBIEN	[1 2 3 4 5]	281.55	142.49	101.39	77.40	59.29	39.16	32.52	26.18	22.59	18.63	17.04	14.27	12.37	12.53	
	(122, 15)	t_p08a14an_2010	[1 2 4 3 5]	302.80	135.17	102.32	75.37	59.15	39.17	32.02	26.15	22.20	20.30	17.63	14.53	13.62	13.33	
	(122, 14)	d_PSINDER	[1 2 4 6 3]	348.41	129.46	102.21	72.66	58.90	44.93	36.25	27.02	24.02	19.72	17.06	14.93	14.23	14.33	
	(122, 13)	d_P15YM_AN	[1 2 4 6 3]	364.28	119.74	104.57	71.21	59.20	43.49	34.58	30.99	26.88	22.70	18.71	16.50	16.26	14.92	
	(122, 12)	t_VPH_snbién_2010	[1 2 4 6 3]	345.53	117.45	105.00	75.12	60.50	42.85	35.83	31.87	26.80	22.30	19.18	17.86	16.95	14.90	
	(122, 11)	t_pander_2010	[1 3 2 5 6]	360.46	109.34	105.64	81.91	60.91	49.02	39.57	32.74	27.20	22.08	20.71	19.20	17.33	15.30	
	(122, 10)	d_VPH_PISOTI	[1 3 4 6 2]	370.71	105.72	111.66	77.80	66.69	47.54	44.92	34.95	30.37	25.27	22.82	20.08	16.52	15.62	
	(122, 9)	prom_hnv_2010	[1 4 3 6 5]	304.87	110.98	117.44	82.10	71.11	47.13	39.89	33.79	29.32	23.95	22.23	20.79	17.83	16.18	
	(122, 8)	t_pdesocup_2010	[1 4 5 2 3]	312.43	131.92	120.10	91.62	65.52	56.15	46.00	37.67	33.78	26.91	22.76	21.17	19.20	16.60	
	(122, 7)	d_PDESOCUP	[1 5 4 2 3]	467.04	242.74	139.86	110.94	78.97	54.40	53.10	44.22	39.41	29.17	26.48	21.50	18.34	17.85	
	(122, 6)	draproces_2010	[1 4 5 2 9]	424.88	222.79	156.21	110.31	82.45	65.02	57.05	48.99	37.64	32.50	27.28	22.87	19.43	17.25	
Davies Bouldin	(122, 22)	Ninguna	[1 2 3 5 6]	0.59	0.75	0.80	0.84	0.87	1.04	1.19	1.22	1.84	1.35	1.31	1.34	1.40	1.47	
	(122, 21)	d_GRAPROES	[1 2 3 5 6]	0.59	0.72	0.78	0.80	0.85	1.00	1.06	1.19	1.34	1.32	1.34	1.41	1.41	1.48	
	(122, 20)	d_PROM_HNV	[1 2 3 5 6]	0.59	0.71	0.77	0.82	0.86	1.00	1.05	1.16	1.26	1.29	1.37	1.43	1.44		
	(122, 19)	d_P15YM_SE	[1 2 3 5 6]	0.60	0.72	0.78	0.82	0.84	0.97	1.01	1.20	1.27	1.32	1.33	1.37	1.41		
	(122, 18)	d_PNACOE	[1 2 3 5 6]	0.55	0.70	0.77	0.82	0.84	1.01	1.06	1.14	1.25	1.31	1.29	1.46	1.41	1.42	
	(122, 17)	d_PBA14AN	[1 2 3 5 6]	0.55	0.70	0.76	0.81	0.84	1.01	1.05	1.14	1.17	1.21	1.30	1.37	1.37	1.32	
	(122, 16)	d_VPH_SNBIEN	[1 2 3 4 5]	0.53	0.71	0.75	0.80	0.84	1.01	1.04	1.10	1.11	1.18	1.19	1.29	1.31	1.26	
	(122, 15)	t_p08a14an_2010	[1 2 4 3 5]	0.51	0.73	0.75	0.81	0.83	1.02	1.03	1.07	1.09	1.13	1.15	1.23	1.27	1.25	
	(122, 14)	d_PSINDER	[1 2 4 6 3]	0.49	0.74	0.78	0.82	0.82	0.92	0.97	1.08	1.08	1.17	1.17	1.32	1.26	1.16	
	(122, 13)	d_P15YM_AN	[1 2 4 6 3]	0.49	0.78	0.76	0.83	0.84	0.94	1.00	0.97	1.01	1.03	1.02	1.19	1.13	1.19	
	(122, 12)	t_VPH_snbién_2010	[1 2 4 6 3]	0.50	0.77	0.77	0.80	0.83	0.94	0.98	0.96	1.01	1.13	1.10	1.16	1.09	1.15	
	(122, 11)	t_pander_2010	[1 3 2 5 6]	0.51	0.81	0.76	0.77	0.83	0.88	0.89	0.93	1.01	1.07	1.02	1.07	1.07	1.14	
	(122, 10)	d_VPH_PISOTI	[1 3 4 6 2]	0.51	0.82	0.73	0.81	0.78	0.84	0.84	0.93	0.94	0.97	1.01	1.06	1.26	1.11	
	(122, 9)	prom_hnv_2010	[1 4 3 6 5]	0.55	0.80	0.71	0.76	0.77	0.88	0.90	0.95	0.97	1.06	1.04	1.07	1.09	1.08	
	(122, 8)	t_pdesocup_2010	[1 4 5 2 3]	0.54	0.74	0.71	0.74	0.86	0.81	0.85	0.89	0.88	0.93	1.02	1.01	1.02	1.03	
	(122, 7)	d_PDESOCUP	[1 5 4 2 3]	0.31	0.54	0.65	0.67	0.73	0.83	0.76	0.82	0.81	0.92	0.94	0.99	1.05	1.03	
	(122, 6)	draproces_2010	[1 4 5 2 9]	0.34	0.44	0.61	0.69	0.79	0.78	0.77	0.78	0.86	0.88	0.95	0.96	1.02	1.09	

Figura 4.26: Resultados *clustering* eliminando las variables de menor relevancia Base CPV 122 localidades Nayarit estandarizada.

En el *embedding* espectral de la Figura 4.27 se observa que las 5 dimensiones tienen un efecto discriminante en las instancias, por lo que se considera apropiado incluir las 5 dimensiones, no obstante se analizan los clústeres con $k = 5$, $k = 4$ y $k = 3$, y se elige aquél que tenga el mejor equilibrio entre interpretabilidad y ajuste.

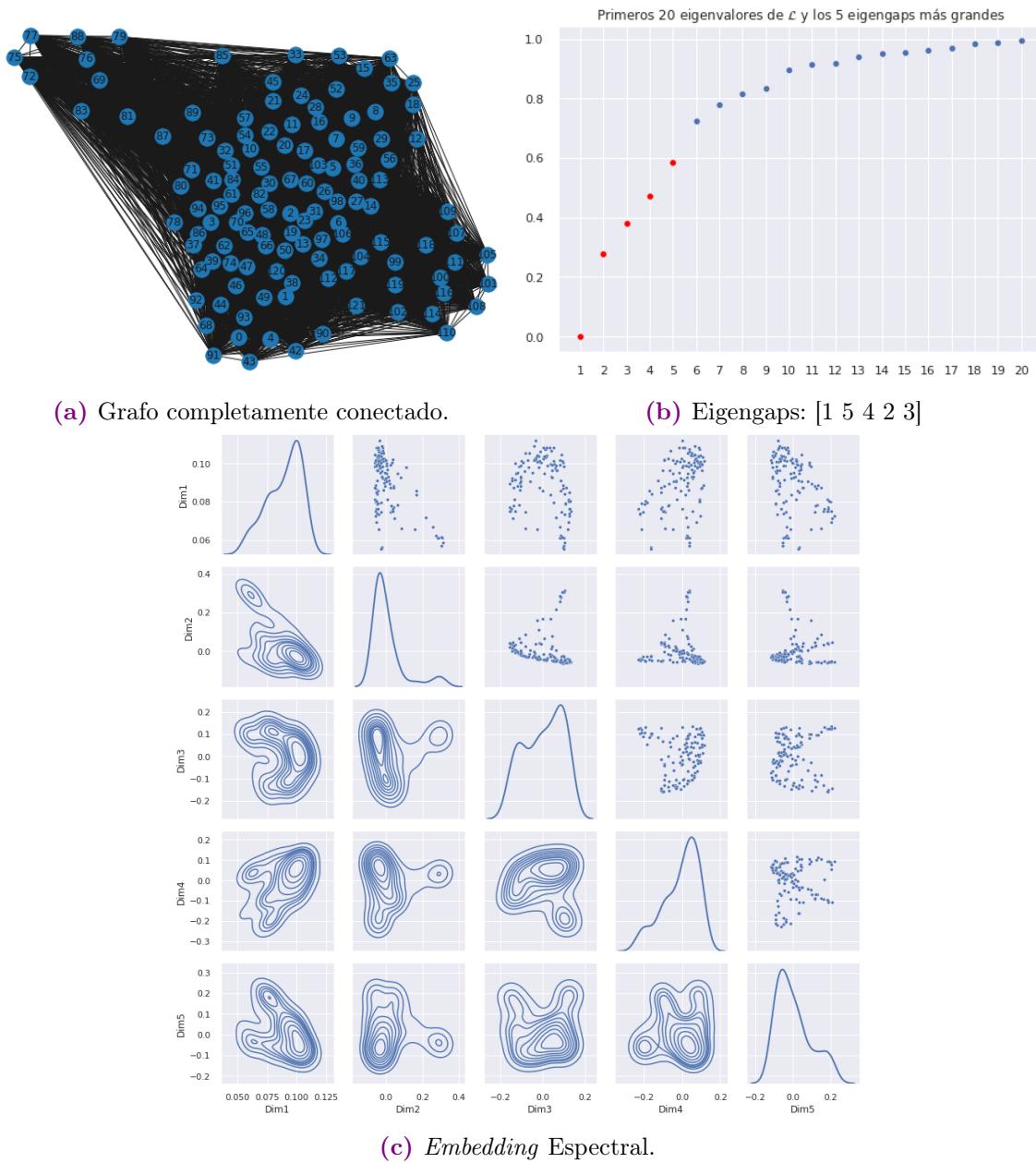


Figura 4.27: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base de 7 variables CPV, 122 localidades Nayarit estandarizada.

Se analizaron las 3 propuestas de *clustering*, con $k = 5$ hay grupos que son muy similares y que se pueden agrupar en uno solo. Por otro lado, la agrupación obtenida con $k = 3$, aunque tiene un mejor ajuste en términos de distorsión, es difícil de interpretar ya que los grupos presentan mucha dispersión. Por lo anterior, se decide que la mejor opción de equilibrio entre ajuste e interpretación es con $k = 4$.

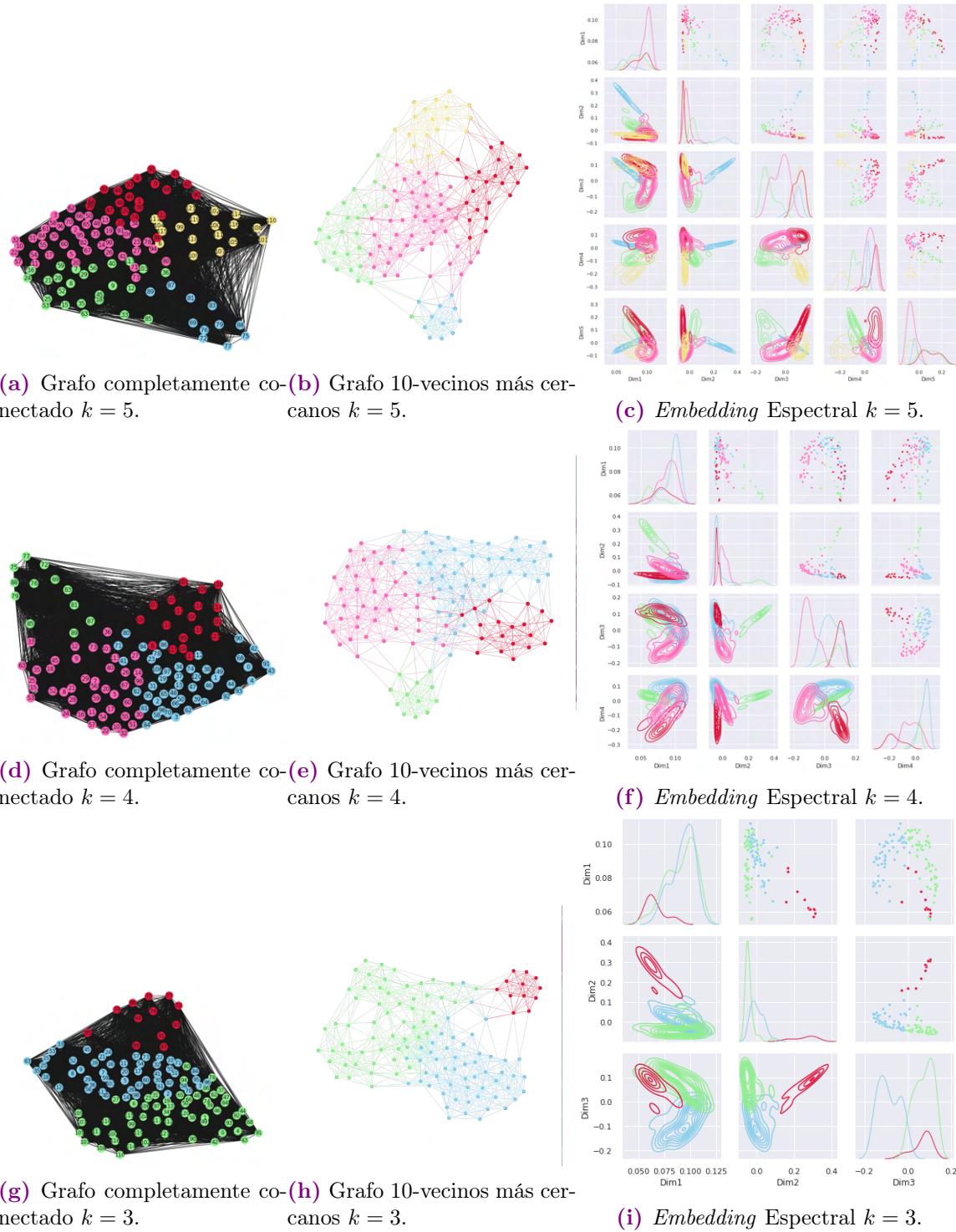


Figura 4.28: Resultados *clustering* espectral con parámetro local σ_i . Base de 7 variables del CPV y 122 localidades Nayarit estandarizada.

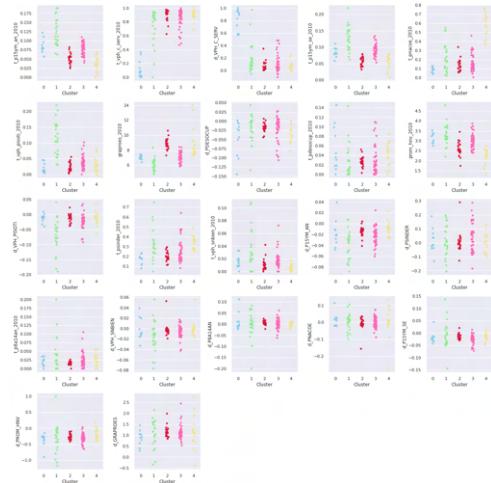
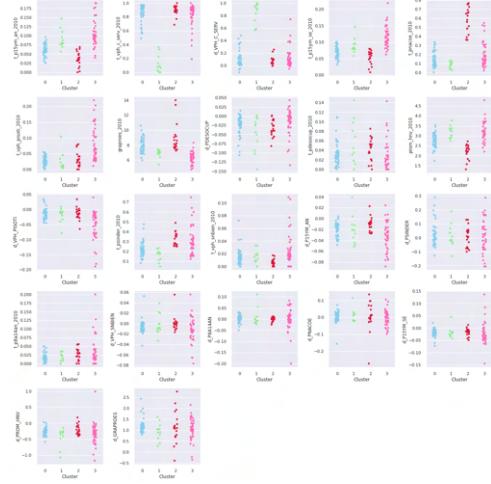
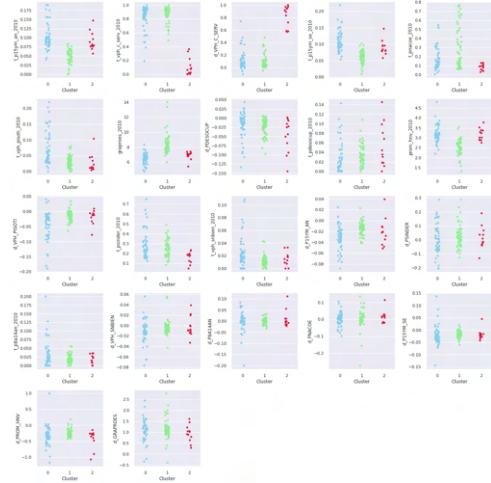
(a) Strip plot $k = 5$.(b) Strip plot $k = 4$.(c) Strip plot $k = 3$.

Figura 4.29: Distribución de clústeres por variable según *ranking* de relevancia. Base de 22 variables del CPV y 122 localidades Nayarit estandarizada.

Eigensearch o Búsqueda iterativa del Eigengap. Clústeres CPV.

Después de haber elegido como versión final la agrupación con $k = 4$ y 7 variables del CPV, aplicamos el algoritmo Eigensearch sobre cada uno de los clústeres generados. En las siguientes gráficas se observa que el gap en 1 es considerablemente mayor al resto de gaps, lo cual indica que ya no hay más clústeres por modelar.

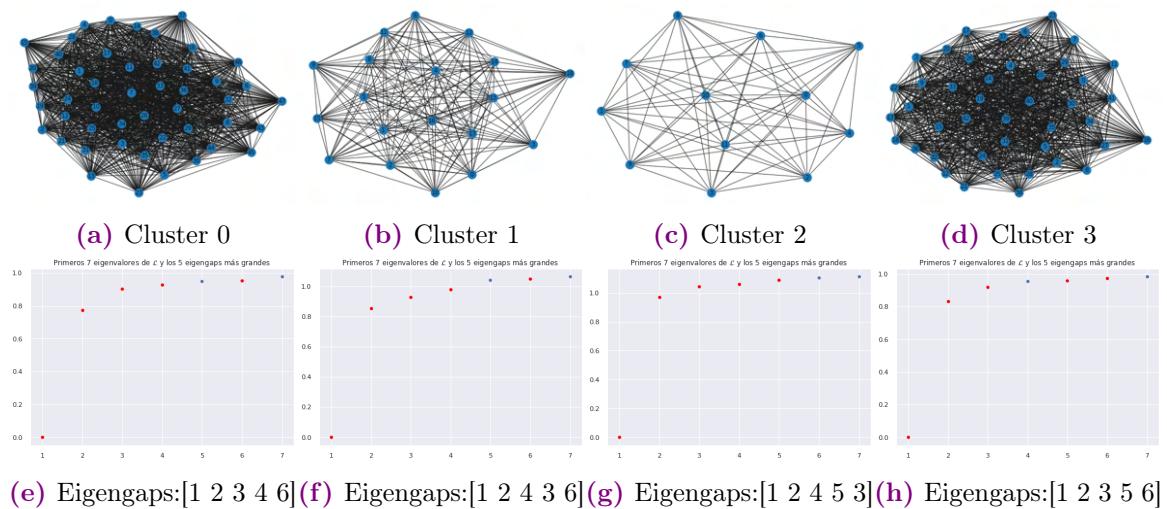


Figura 4.30: Eigensearch.

Interpretación Clústeres CPV.

Los 4 clústeres generados con las variables del CPV para el estado de Nayarit quedan distribuidos de la siguiente manera:

Cluster	Localidades	%
Azul	48	39 %
Rosa	43	35 %
Rojo	19	16 %
Verde	12	10 %
Total	122	100 %

Tabla 4.4: Clústeres CPV Nayarit.

La interpretación de los grupos se basa principalmente en la dispersión de las 7 variables de mayor relevancia y sus cuantiles poblacionales. Asimismo, se visualiza el resumen de dicha información a través del *embedding* espectral.

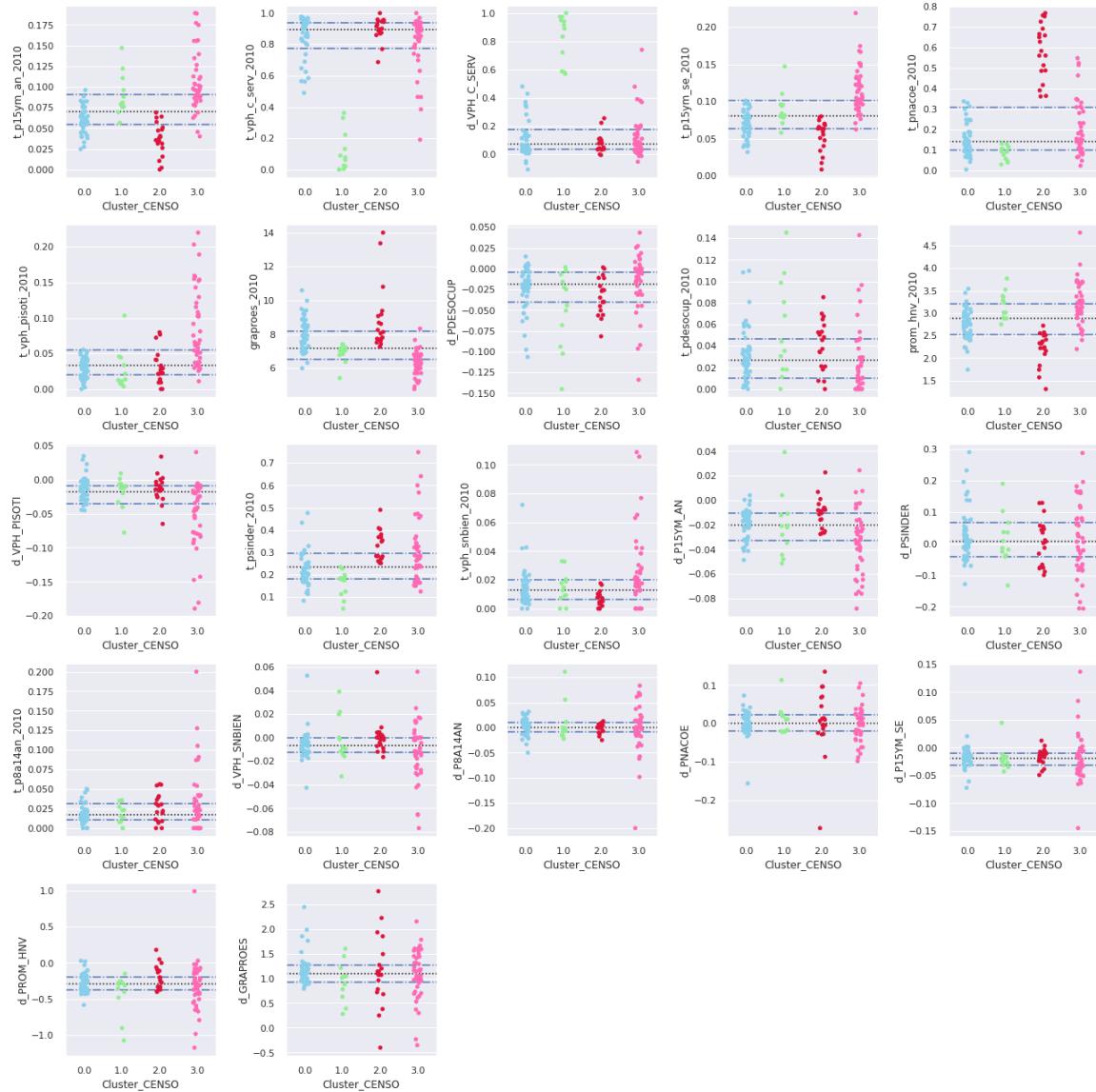


Figura 4.31: Dispersion clústeres CPV y cuantiles .25, .50 y .75 poblacionales.

En la Figura 4.32 se muestra la correlación lineal entre las variables y las dimensiones del *embedding*.

La dimensión 1 se relaciona con la cobertura de servicios básicos en la localidad, sin embargo, todas las observaciones se encuentran distribuidas uniformemente dentro de un rango, por lo tanto, esta dimensión no provee información significativa. Esto se debe a que la primera dimensión corresponde al eigenvector del primer eigenvalor, cuyo valor siempre es 0.

La dimensión 2 esta fuertemente relacionada positivamente a localidades que con-

taban con una cobertura baja de servicios básicos y altos niveles de personas sin escolaridad o analfabetas en 2010, no obstante, esto ha mejorado en los últimos años. Estas localidades se encuentran en el clúster verde.

La dimensión 3 distingue a las localidades que desde 2010 contaban con mejores condiciones ya que se relaciona positivamente con un grado de escolaridad alto y un mayor porcentaje de habitantes nacidos en otra entidad. En contra parte, valores más negativos se relacionan a localidades con mayor analfabetismo, sin escolaridad y viviendas con piso de tierra.

La dimensión 4 se relaciona fuertemente con el porcentaje de habitantes nacidos en otra entidad.

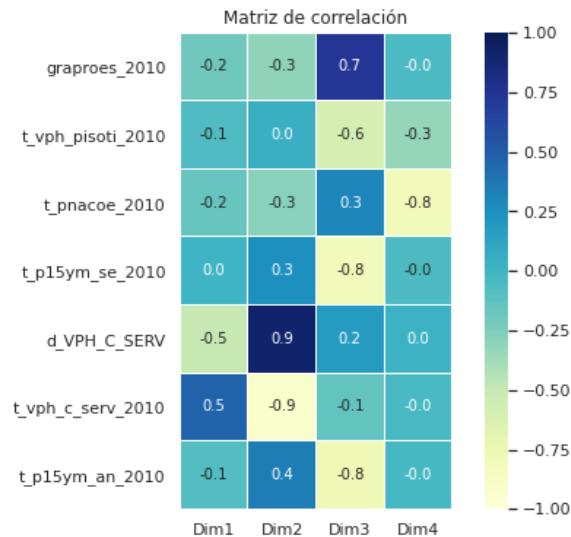


Figura 4.32: Matriz de correlación variables originales vs. *embedding*.

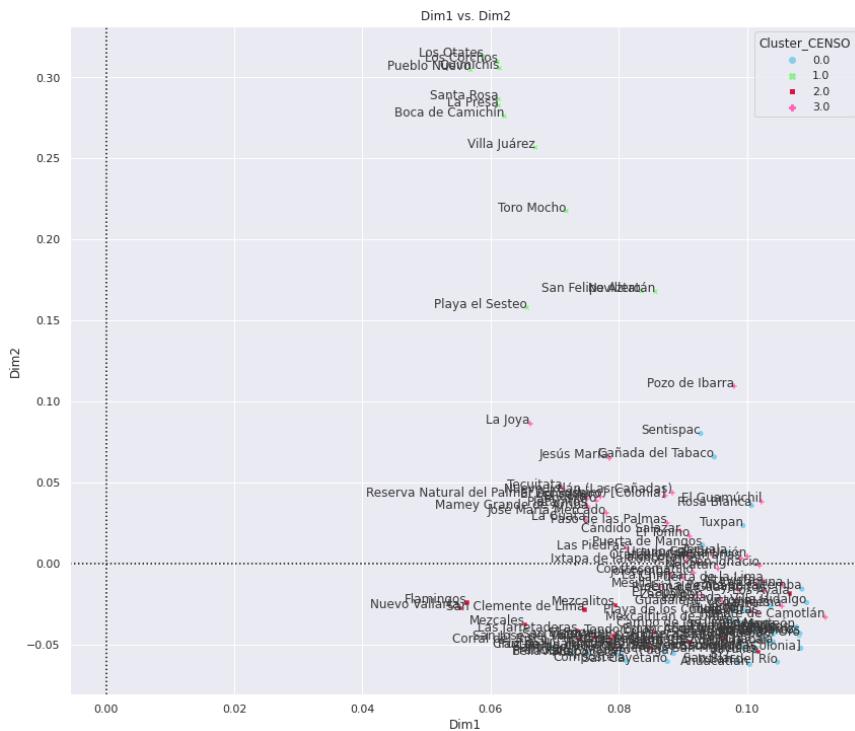


Figura 4.33: Dim1 vs. Dim2

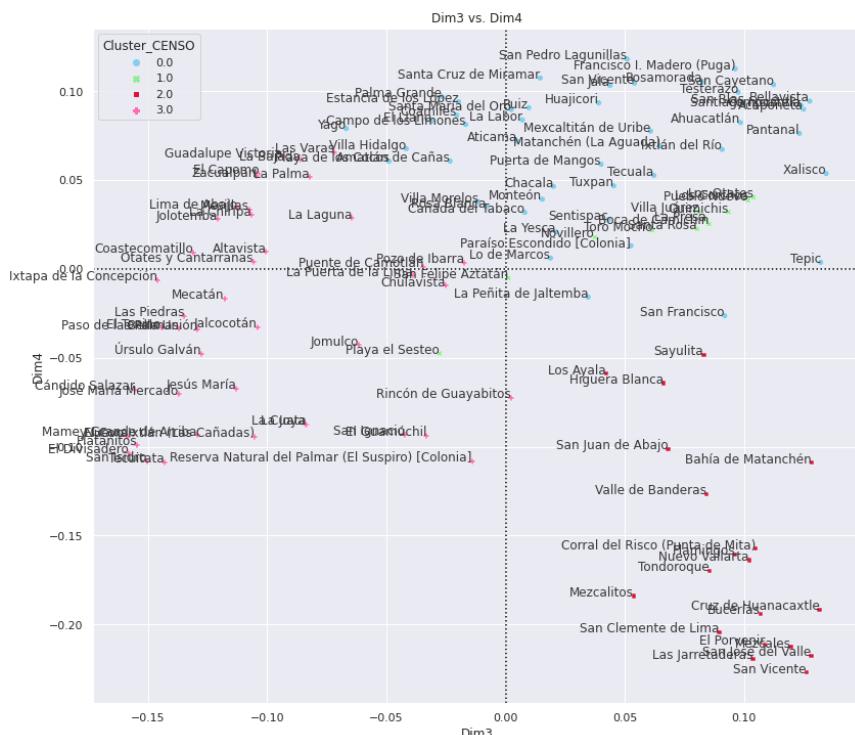


Figura 4.34: Dim3 vs. Dim4

Clúster Azul:

- Se caracteriza por ser el clúster que se encuentra dentro del nivel promedio en las variables referentes al porcentaje de analfabetismo, de viviendas que cuentan con servicios básicos, de personas de 15 años y más sin escolaridad, de población nacida en otra entidad y de viviendas habitadas con piso de tierra.
- Es el segundo clúster con mayor grado de escolaridad, después del clúster rojo. La mitad de las localidades en este clúster tienen un grado mayor a 7.9 años.
- En este clúster se encuentra Tepic, Santiago Ixcuintla, San Blas, Matachén, Jalisco, San Francisco, Compostela, etc.

Todo lo anterior sugiere que las localidades que lo conforman no sobresalen por ser mejor o peor, simplemente tienen un nivel medio con respecto a todas las localidades en Nayarit.

Clúster Verde:

- Este clúster tiene un nivel alto de analfabetismo y en 2010 registraban un porcentaje bajo de viviendas con servicios básicos, sin embargo, en 2020 se observa una mejora importante en este aspecto.
- Registra porcentajes bajos de personas nacidas en otra entidad y un bajo grado de escolaridad después del clúster rosa. El 75 % de las localidades en este clúster tienen un grado menor a 7.2 años.
- En este clúster se encuentra Playa Sesteo, ToroMocho, Santa Rosa, entre otros.

Clúster Rojo:

- Este clúster presenta los mejores indicadores de bienestar social, presenta los niveles más bajos de analfabetismo y personas sin escolaridad.
- En estas localidades una gran parte de los habitantes nacieron en otra entidad, en 75 % de las localidades 49 % de los habitantes nacieron en otra entidad

- El grado de escolaridad es de los más altos y número promedio de hijos es el más bajo. La mitad de las localidades tienen más de 8.1 años y las localidades con mayor grado son Flamingos, Nuevo Vallarta, Mezcales, San Clemente de Lima, Bahía de Matachén, Cruz de Huanacaxtle, San José del Valle y San Vicente.

Clúster Rosa:

- Este clúster presenta los peores indicadores de bienestar social, presenta los niveles más altos de analfabetismo, personas sin escolaridad y de número de hijos.
- Presenta las tasas más altas de viviendas con piso de tierra, aunque en los últimos años ha disminuido.
- El grado de escolaridad es de los más bajos. La mitad de las localidades tienen menos de 6.4 años, las de menor grado son Mamey Grande de Arriba, Platanitos, El Divisadero, entre otros.

Con base en el análisis anterior se propone el siguiente *ranking* por nivel de bienestar, donde 4 estrellas es el máximo nivel y una estrella en nivel mínimo:

Cluster CPV	Nivel de Bienestar	Categoría
Rojo	★★★	A
Azul	★★	B
Verde	★	C
Rosa	*	D

Tabla 4.5: Ranking Clústeres CPV.

Una vez obtenido el *ranking* tanto del conjunto de variables del DENUE como del CPV es posible identificar qué localidades presentan un escenario favorecedor para los micro negocios y además cuentan con los mejores niveles de bienestar en la entidad. Dichas localidades se encuentran rankeadas de acuerdo a la suma total de estrellas, por lo que el mejor clúster será la categoría A-A de 8 estrellas en total y posteriormente la categoría A-B o B-A con 7 estrellas, y así sucesivamente.

En el estado de Nayarit, 2 localidades son categoría A-A, Los Ayala y Nuevo Vallarta, y 21 localidades son categoría A-B/B-A, las cuales refieren a las localidades

		Clúster CENSO							
		.	0	-1	A	B	C	D	NA
Clúster DENUÉ	0	224							
	-1		2500						
	A			2	6	1	8	4	
	B			15	38	9	27		
	C			2	4	2	8		

Figura 4.35: Localidades Nayarit por categoría.

que presentan potencial en el sector micro pero que las condiciones de bienestar social se han mantenido en niveles normales respecto al resto de localidades, o al contrario, presentan buen nivel de bienestar social pero niveles normales de crecimiento del sector micro. En la figura 4.36, se puede observar que estas localidades se encuentran ubicadas al sur de estado, en la costa y muy cerca de la frontera con Jalisco. El municipio que concentra el mayor número de localidades potenciales es Bahía de Banderas (17) y el resto en San Blas (2), Compostela (2) y Santiago Ixcuintla (2).

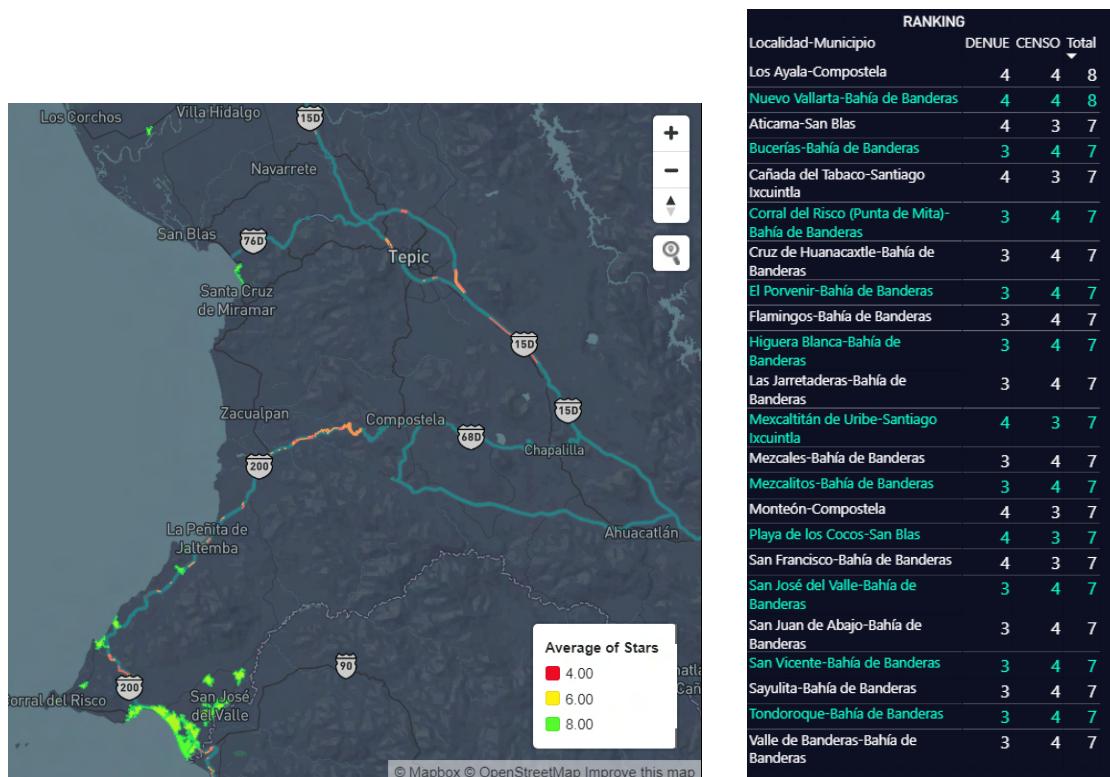


Figura 4.36: Localidades con mayor potencial de crecimiento y bienestar social Nayarit.

Cabe mencionar que en Nuevo Vallarta, predomina el sector *Comercio al por*

menor, en particular el comercio de productos textiles, bisutería, accesorios de vestir y de calzado, así como los *Servicios de preparación de alimentos y bebidas*. También predominan los servicios de banca múltiple, sin embargo, aunque son considerados micro negocios por su tamaño de personal, realmente estos pertenecen a grandes corporaciones.

Por otra parte, en la localidad Los Ayala predomina fuertemente los *Servicios de alojamiento temporal* (cabañas, villas y similares) y los *Servicios de preparación de alimentos y bebidas alcohólicas*.

Finalmente, tomando como base las 122 localidades que pertenecen a alguna de las categorías (A,B,C,D) se obtuvo el siguiente mapa a nivel municipio iluminado de acuerdo al promedio de puntaje de las localidades que lo constituyen. En primer lugar se ubica Bahía de Banderas y en último lugar Del Nayar.

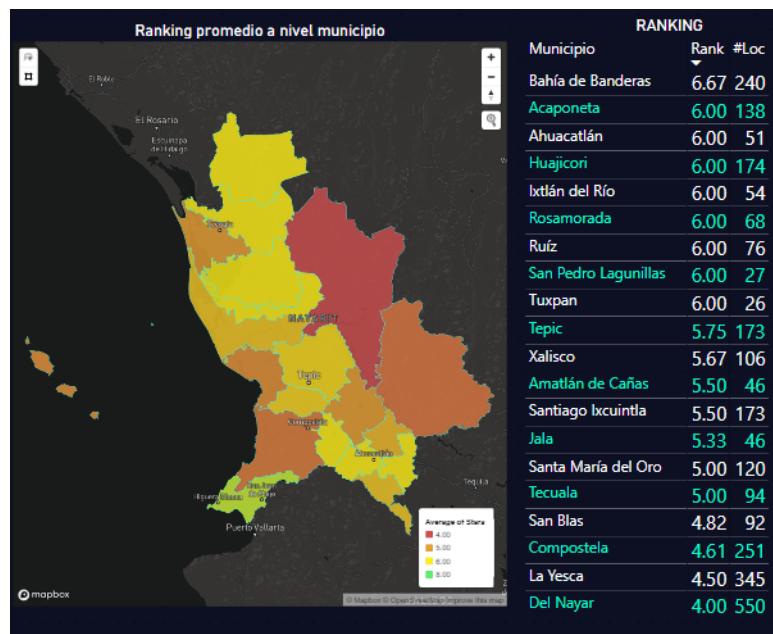


Figura 4.37: Ranking por municipio Nayarit.

4.2. Nuevo León

4.2.1. Análisis exploratorio

El estado de Nuevo León está constituido por 51 municipios y 4822 localidades con más de 1 o 2 viviendas, el 79 % son rurales puntuales, 19 % rurales y 2 % urbanas. Gran parte sus localidades (93 %) no tienen ningún registro de UE Micro en el DENUE, por lo tanto se colocan dentro del *Cluster -1* de nula actividad económica del sector Micro.

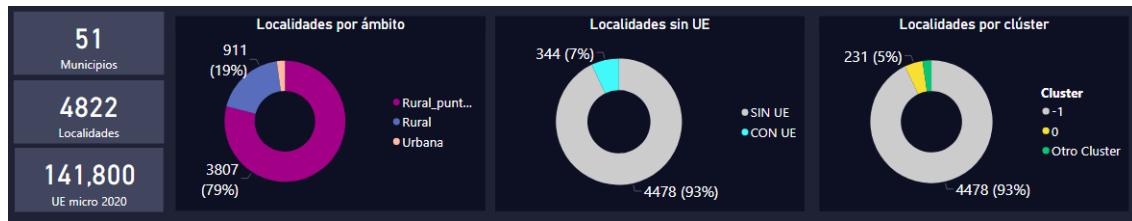


Figura 4.38: Información general Nuevo León

Las 344 localidades restantes que si cuentan con UE se localizan en el centro del estado ([4.66](#)).

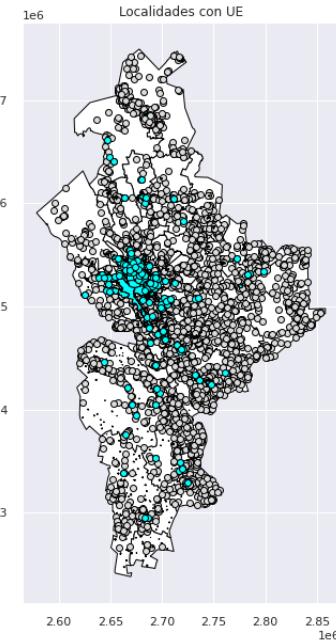


Figura 4.39: Localidades con Unidades Económicas

Sin embargo, cabe resaltar que un subconjunto de estas 344 localidades han registrado 5 o menos UE activas activas en 2020, estas localidades se agrupan en el *Cluster 0* de baja actividad económica micro.

A continuación se muestran las localidades que quedaron agrupadas en el *Cluster -1* (gris), *Cluster 0* (amarillo) y Otros(verde).

Cluster	Descripción	Localidades	Urbanas	Rurales	Rurales puntuales
-1	Nula actividad micro	4,478	0	809	3,669
0	Baja actividad micro	231	10	86	135
Otros	Loc. por clusterizar	113	94	16	3

Tabla 4.6: Distribución localidades Nuevo León. Cluster -1, 0 y Otros.

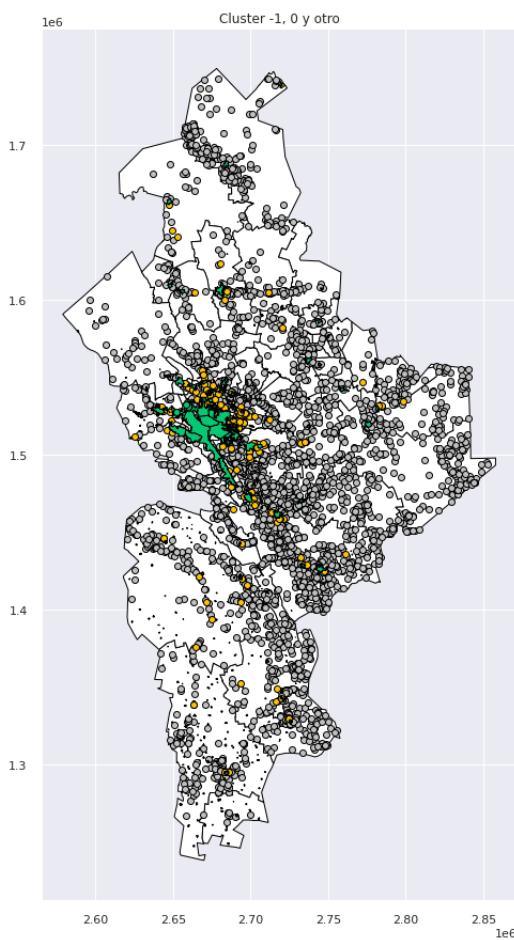


Figura 4.40: Distribución localidades Nuevo León. Cluster -1, 0 y Otros.

4.2.2. Selección de variables mediante algoritmo SPEC.

Considerando la base de datos de las 113 localidades a clusterizar, se aplica el algoritmo SPEC para identificar las variables que tienen mayor influencia en la formación de los clústeres. Los parámetros utilizados y función de rankeo son los mismos que se utilizaron para los datos de Nayarit, por lo tanto valores pequeños en el score $\varphi_2(F_i)$ indican que la variable F_i se alinea estrechamente con los eigenvectores no triviales de los eigenvalores más pequeños y por lo tanto la i -ésima variable provee una buena separabilidad de las observaciones.

Al igual que para los datos de Nayarit, al rankear el conjunto de variables del DENUE y del CPV por separado se obtuvieron mejores resultados que uniendo ambos conjuntos de variables. El *ranking* para cada conjunto se muestra a continuación. Cabe mencionar que de las 113 localidades a clusterizar, 15 no cuentan con información completa en las variables del CPV, por lo tanto, en este grupo de variables se consideran 98 localidades por agrupar.

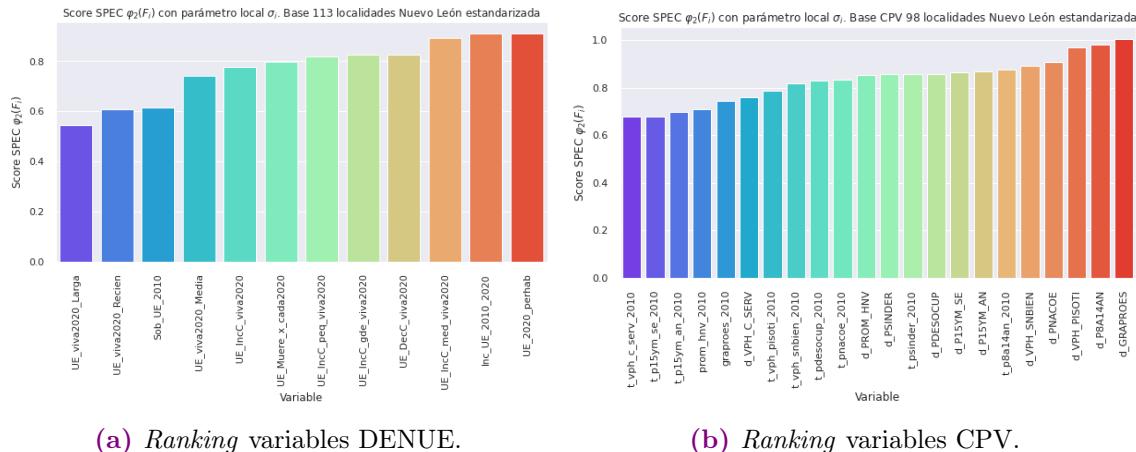


Figura 4.41: Resultados algoritmo SPEC con parámetro local σ_i . Base DENUE y base CPV Nuevo León estandarizada.

En la figura 4.41a se observa que las variables que más influyen en la clusterización son las referentes a la antigüedad de las UE y las de menor relevancia son las variables que miden el número de UE por habitante y la tasa de UE que pasaron de ser micro a medianas empresas. Por otra parte, en la figura 4.41b se aprecia que las primeras 8 o 9 variables del CPV son las mayor relevancia, y que a partir de éstas el score $\varphi_2(F_i)$

se mantiene casi constante.

4.2.3. *Clustering* Espectral.

La metodología y lógica que se utiliza para encontrar la estructura final de los clústeres es similar a la aplicada en los datos de Nayarit, es decir, se obtienen clústeres por separado para las variables del DENUE y del CPV, para cada caso primero se obtienen las agrupaciones considerando todas las variables y en caso de no encontrar un buen ajuste e interpretabilidad se eliminan las variables de menor relevancia hasta encontrar la mejor opción de *clustering*.

A continuación se describen los resultados obtenidos.

Grafo completamente conectado, 12 variables DENUE.

Al considerar todas las variables del DENUE, el grafo de similitudes no parece revelar grupos de una manera clara y esto se ve reflejado en los eigengaps, la brecha más grande la observamos en el primer eigenvalor lo cual indica que el número de clústeres sea igual a 1. Por otra parte, en el *embedding* se logran identificar a partir de las curvas de nivel a lo más 4 clústeres.

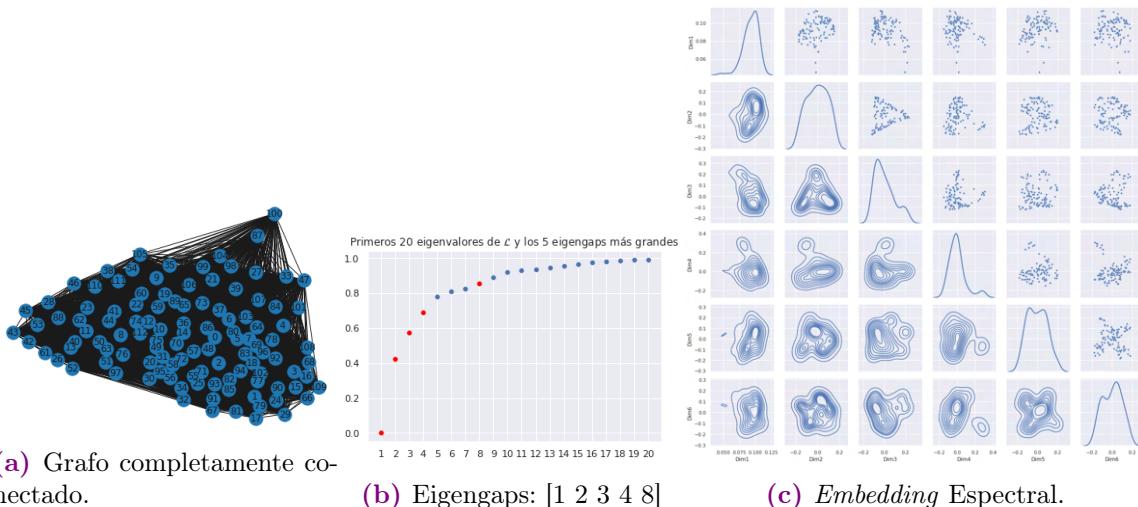


Figura 4.42: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base de 12 variables DENUE, 113 localidades Nuevo León estandarizada.

Con base en lo anterior, se realizaron diversas pruebas considerando distinto nú-

mero de clústeres, sin embargo, los grupos formados tienen mucha distorsión o en el caso de $k = 2$ se tienen clústeres muy generalizados que no revelan información interesante.

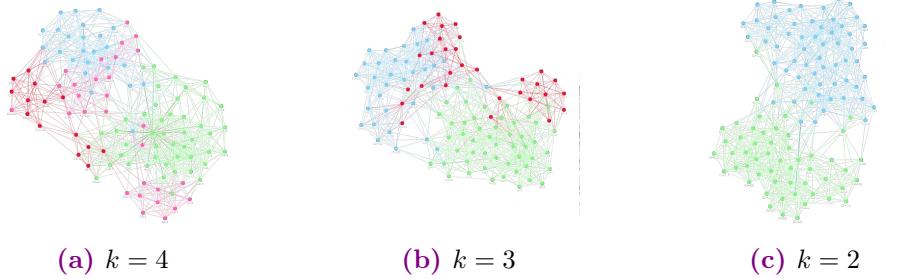


Figura 4.43: clustering $k = 4$, $k = 3$ y $k = 2$. Base de 12 variables DENUE, 113 localidades Nuevo León estandarizada.

Por lo anterior, se decide verificar si al eliminar las variables de menor relevancia se obtiene un grafo más claro y en consecuencia clústeres más claros.

Grafo completamente conectado, eliminando variables de menor relevancia DENUE.

Se realiza el ejercicio de eliminar de una en una las variables de menor relevancia.

En la Figura 4.44 se observa una mejora importante en el estadístico Silhouette y el heurístico eigengap al eliminar las 3 variables de menor relevancia.

Además, el segundo gap más grande pasa de 2 a 4 indicando que posiblemente haya 4 clústeres en el grafo y el criterio Silhouette mejora en 4 y 7 décimas al elegir $k = 3$ y $k = 4$ clústeres respectivamente.

Valores	Dim	V. Eliminada	Eigengap Top5	Número de Clusters														
				2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Silhouette	(113, 12)	Ninguna	[1 2 3 4 8]	0.59	0.50	0.46	0.39	0.41	0.42	0.47	0.44	0.39	0.36	0.35	0.33	0.34	0.32	
	(113, 11)	UE_2020_perhab	[1 2 3 4 5]	0.59	0.50	0.46	0.40	0.40	0.43	0.47	0.44	0.40	0.38	0.37	0.34	0.33	0.33	
	(113, 10)	Inc_UE_2010_2020	[1 2 4 3 6]	0.60	0.50	0.50	0.42	0.41	0.44	0.47	0.43	0.40	0.38	0.37	0.34	0.33	0.31	
	(113, 9)	UE_IncC_med_viva202	[1 4 2 6 5]	0.60	0.54	0.53	0.46	0.44	0.46	0.47	0.43	0.39	0.37	0.37	0.36	0.33	0.33	
	(113, 8)	UE_DecC_viva2020	[1 2 4 8 6]	0.63	0.57	0.52	0.52	0.46	0.46	0.46	0.42	0.39	0.38	0.35	0.34	0.34	0.35	
	(113, 7)	UE_IncC_gde_viva202	[1 2 4 3 8]	0.66	0.59	0.53	0.47	0.45	0.43	0.42	0.39	0.37	0.35	0.35	0.34	0.33	0.32	
	(113, 6)	UE_IncC_peq_viva202	[1 2 4 6 3]	0.68	0.62	0.56	0.51	0.51	0.42	0.39	0.37	0.35	0.34	0.34	0.35	0.35	0.36	
	(113, 5)	UE_Muere_x_cada202	[2 1 3 7 6]	0.71	0.62	0.58	0.55	0.47	0.48	0.50	0.44	0.42	0.41	0.38	0.39	0.39	0.40	
	(113, 4)	UE_IncC_viva2020	[3 6 2 1 7]	0.72	0.63	0.62	0.54	0.50	0.51	0.51	0.48	0.43	0.41	0.40	0.42	0.37	0.36	
	(113, 3)	UE_viva2020_Media	[5 2 3 7 9]	0.67	0.63	0.59	0.58	0.53	0.53	0.49	0.52	0.50	0.47	0.43	0.46	0.45	0.43	
	(113, 2)	Sob_UE_2010	[6 3 12 8 10]	0.71	0.69	0.51	0.54	0.57	0.58	0.57	0.53	0.53	0.52	0.57	0.52	0.49	0.46	
	(113, 1)	UE_viva2020_Recien	[9 8 13 7 11]	0.97	0.75	0.71	0.72	0.73	0.77	0.80	0.75	0.74	0.72	0.68	0.70	0.67	0.64	

Figura 4.44: Resultados al eliminar las variables de menor relevancia. Base DENUE 113 localidades Nuevo León estandarizada.

Tomando como base las 9 variables más relevantes, se obtiene el grafo de similitudes y su correspondiente *embedding* espectral, en este último se observan 4 poblaciones, por lo que $k = 4$ se perfila a ser el número óptimo de clústeres. No obstante se realizan pruebas con diferente número de clústeres para garantizar buena interpretabilidad, finalmente se elige como el valor óptimo $k = 4$ clústeres.

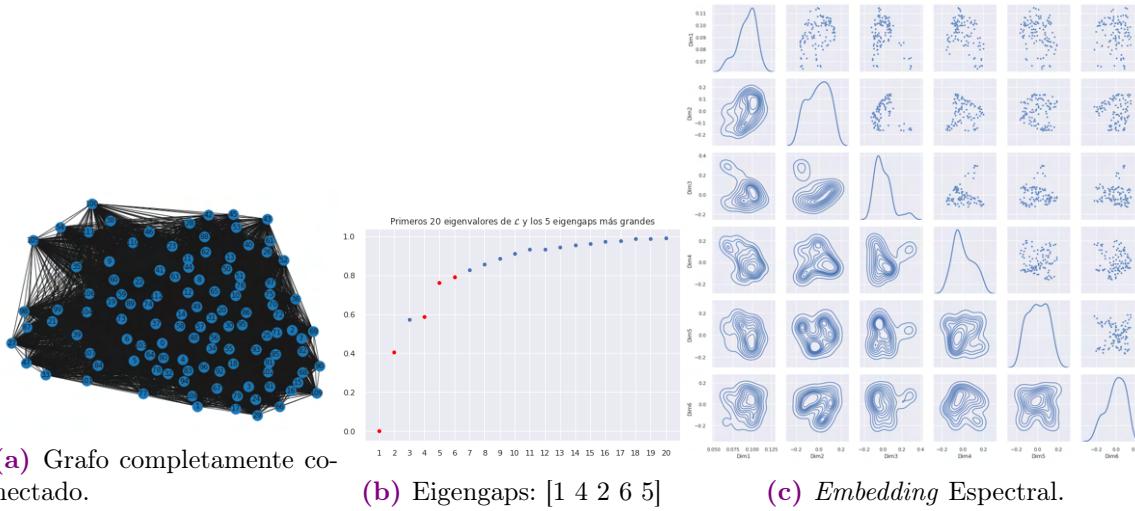


Figura 4.45: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base de 9 variables DENUE, 113 localidades Nuevo León estandarizada.

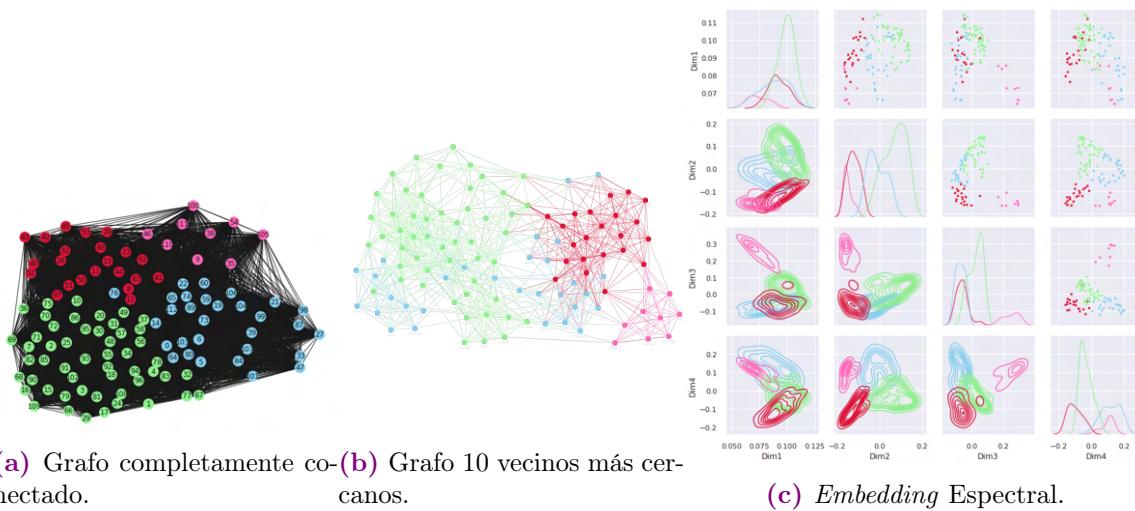


Figura 4.46: Resultados *clustering* espectral $k = 4$ con parámetro local σ_i . Base de 9 variables y 113 localidades Nuevo León estandarizada.

Eigensearch o Búsqueda iterativa del Eigengap. Clústeres DENUE

Después de haber elegido como versión final la agrupación con $k = 4$ y 9 variables del DENUE, verificamos que en cada clúster ya no haya grupos por modelar.

En las gráficas vemos que en todos los casos el primer gap es considerablemente mayor al resto de gaps, lo cual indica que ya no hay grupos por modelar dentro de los clústeres finales.

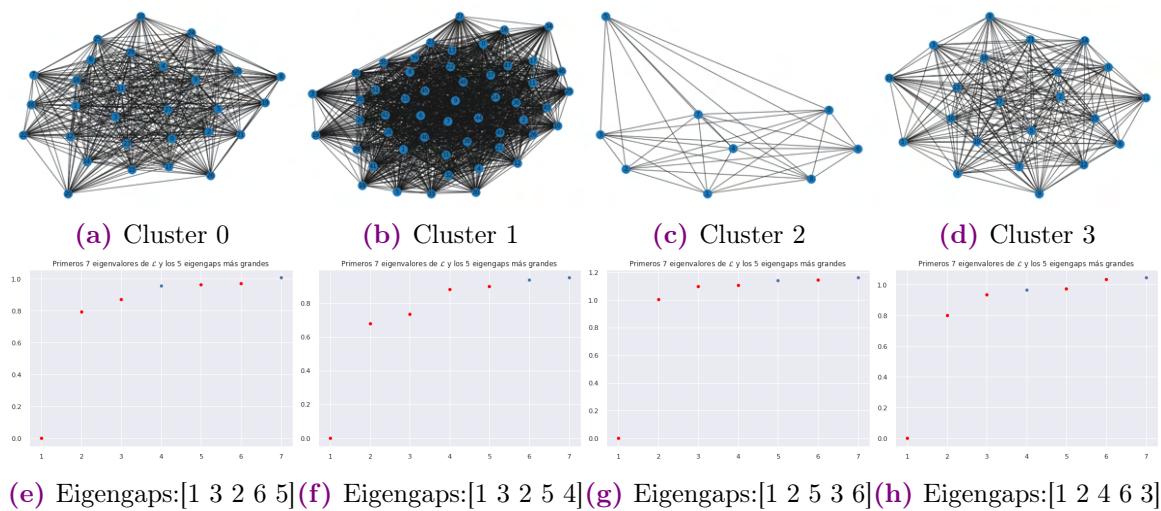


Figura 4.47: Eigensearch.

Interpretación Clústeres DENUE

Los clústeres finales, considerando el conjunto de variables del DENUE, quedaron conformados de la siguiente manera:

Cluster	Localidades	%
Verde	52	46 %
Azul	30	27 %
Rojo	22	19 %
Rosa	9	8 %
Total	113	100 %

Tabla 4.7: Clústeres DENUE Nuevo León.

A continuación se realiza la interpretación de clústeres con base en la dispersión de las variables originales y sus cuatiles poblacionales, además, se muestra cómo a partir

del *embedding* espectral se logra resumir esta información por medio de la distribución de los puntos en el espacio de dimensión reducida.

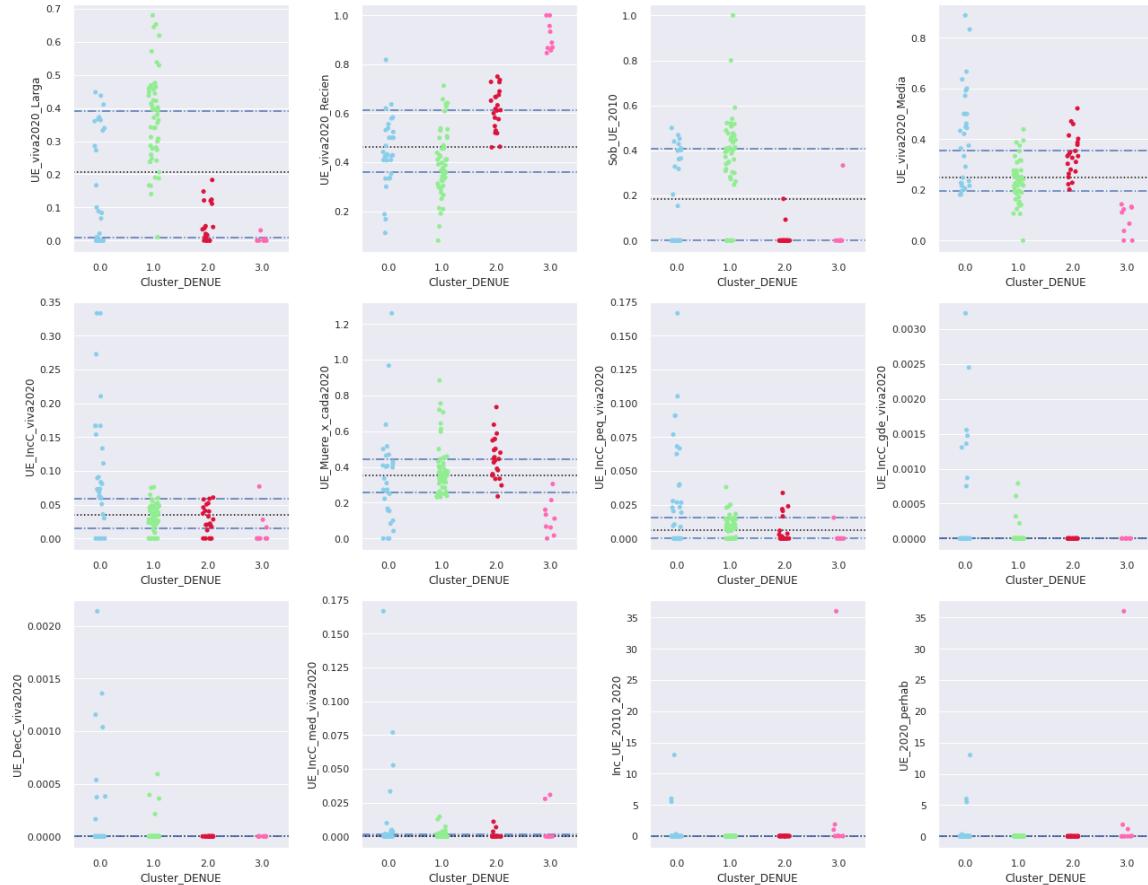


Figura 4.48: Dispersion clústeres DENUE y cuantiles .25, .50 y .75 poblacionales.

En la Figura 4.49 se muestra la correlación lineal entre las variables y las dimensiones del *embedding*.

La dimensión 1 presenta una débil correlación con todas las variables y esto se debe a que corresponde al eigenvector del primer eigenvalor, cuyo valor siempre es 0.

La dimensión 2 esta fuertemente relacionada positivamente a localidades con antigüedades largas cuya tasa de sobre-vivencia es alta y negativamente con antigüedades más bajas.

La dimensión 3 distingue a las localidades de antigüedad media y baja, valores positivos se relacionan con las localidades más jóvenes.

La dimensión 4, se relaciona positivamente a localidades que han presentado cre-

cimientos en tamaño de personal, pasaron de ser micro a pequeñas empresas, por lo que esta última dimensión cobra mayor importancia.

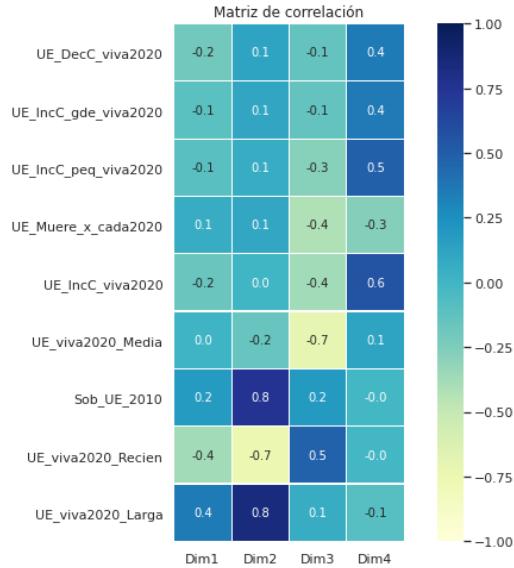


Figura 4.49: Matriz de correlación variables originales vs. *embedding*.



Figura 4.50: Dim1 vs. Dim2

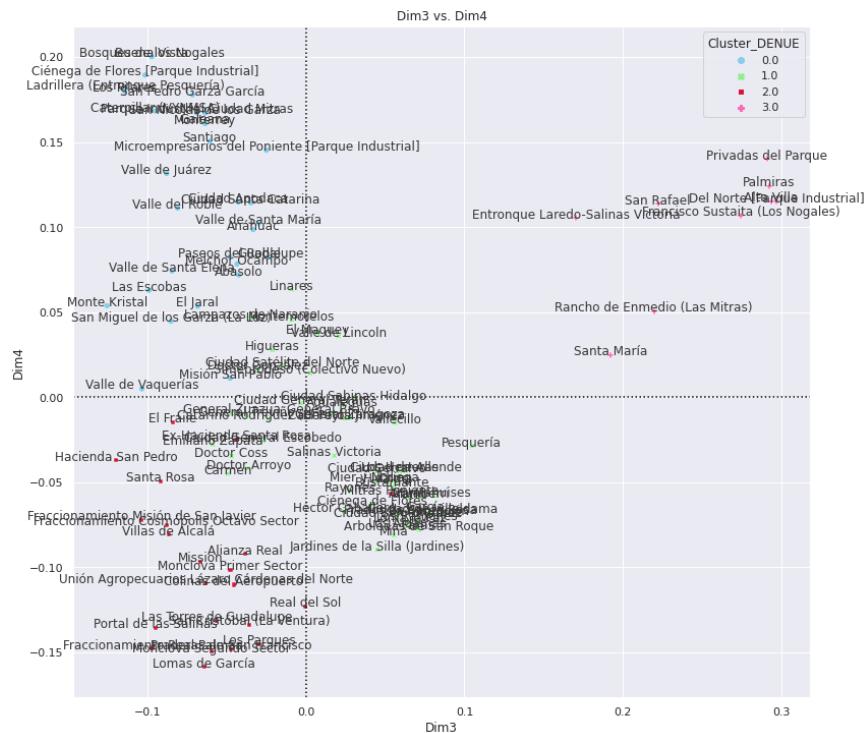


Figura 4.51: Dim3 vs. Dim4

Clúster Verde:

- En este clúster se encuentran las UE con mayor antigüedad, las localidades de este grupo tienen en promedio 37 % de UE dadas de alta antes de 2014 y 40 % de las UE dadas de alta en 2010 han sobrevivido hasta 2020 por lo que tiene una tasa de sobrevivencia a largo plazo alta.
- Tiene los porcentajes más bajos de UE creadas a partir de 2019 (de recién creación) con respecto al resto de localidades, en promedio 40 %.
- Los incrementos en el número personal de las UE se encuentran dentro de los niveles regulares respecto al resto de localidades.

Las localidades en este clúster son de las más consolidadas pero a la vez no registran un crecimiento fuera de los niveles normales. Algunas de las localidades en este clúster son Ciudad General Escobedo, García, Bustamante, China, Iturbide, Montemorelos, Linares, Pesquería, entre otros.

Clúster Azul:

- Este clúster es más jovén que el verde, predominan las UE con antigüedad media, creadas entre 2014 y 2018 (en promedio 40 %). Algunas de las localidades son San Nicolás de los Garza, Galeana, Ciudad de Apodaca, San Pedro Garza García, Santiago, etc.
- En este clúster se encuentran las localidades con mayor número de crecimientos de empresas micro a pequeña, mediana y grande. Destacan Ciénega de Flores, Bosques de los Nogales, Microempresarios del Poniente, San Pedro Garza García, Parque Industrial Ciudad Mitrás, Anáhuac, Ciudad de Apodaca, Caterpillar(VYNMSA), entre otros.
- Las UE que han muerto por cada una de las que sobreviven en 2020 se encuentra en un nivel de normal a bajo respecto al resto de localidades, las localidades que registran niveles muy bajos son Valle de Santa Elena, Paseos del Roble, Valle de Santa María, etc.

Clúster Rojo:

- La mayoría de UE en este grupo presenta antigüedades de media a reciente.
- Una minoría de sus unidades son de largo plazo y presentan una tasa de sobre-vivencia a largo plazo baja.
- Algunos fraccionamientos grandes se encuentran en este clúster.
- Crecimientos de micro a pequeña empresa dentro de los niveles normales
- Es el clúster con más muertes por cada UE viva en 2020.

Clúster Rosa:

- Es el clúster con los negocios más jóvenes, en la mayoría de las localidades el 87 % de las UE se dieron de alta a partir de 2019.

- Se han dado muy pocos saltos de micro a pequeña empresa en localidades como Entronque Laredo-Salinas Victoria y Del Norte [Parque industrial].

Con base en la descripción anterior, se propone el siguiente *ranking* por nivel de potencial del sector micro, donde 4 estrellas es el nivel máximo y una estrella es nivel mínimo:

Cluster DENUE	Potencial sector Micro	Categoría
Azul	★★★	A
Verde	★★	B
Rojo	★	C
Rosa		D

Tabla 4.8: *Ranking* Clústeres DENUE Nuevo León.

Grafo completamente conectado, 22 variables CPV

Se obtiene el grafo de similitudes considerando las 22 variables del CPV, el resultado es una nube de observaciones homogénea en la que aparentemente no existen clústeres. Así mismo, el eigengap más grande se encuentra en el primer eigenvalor indicando que el número de agrupaciones en el grafo es igual a 1. En el *embedding* espectral, se logran apreciar al menos 2 poblaciones en las 3 primeras dimensiones, sin embargo, tampoco es clara la distinción entre ellas.

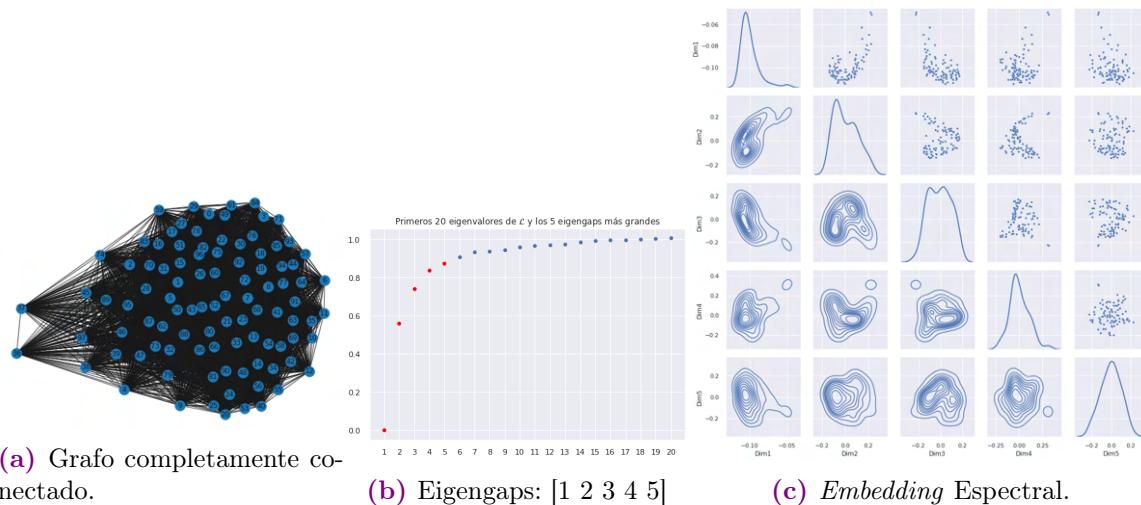


Figura 4.52: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base de 22 variables CPV, 98 localidades Nuevo León estandarizada.

Bajo este escenario, se considera conveniente eliminar las variables de menor relevancia hasta el punto donde los clústeres comiencen a ser más evidentes.

Grafo completamente conectado, eliminando variables de menor relevancia CPV

Se realiza el ejercicio de eliminar de una en una las variables de menor relevancia de acuerdo al *ranking* del algoritmo SPEC. En la Figura 4.53 se observa una mejora importante en el estadístico Silhouette y el heurístico eigengap al eliminar las 9 variables de menor relevancia. Además, el segundo gap más grande pasa de 2 a 3 indicando que posiblemente hay 3 clústeres en el grafo.

Valores	Dim	V. Eliminada	Eigengap Top5	Número de Clusters													
				2	3	4	5	6	7	8	9	10	11	12	13	14	15
Silhouette	(98, 22)	Ninguna	[1 2 3 4 5]	0.64	0.50	0.46	0.36	0.31	0.27	0.26	0.27	0.29	0.28	0.26	0.26	0.26	0.25
	(98, 21)	d_GRAPROES	[1 2 3 4 6]	0.65	0.51	0.46	0.37	0.31	0.29	0.30	0.27	0.28	0.27	0.27	0.25	0.25	0.26
	(98, 20)	d_P8A14AN	[1 2 3 4 6]	0.66	0.51	0.45	0.37	0.31	0.29	0.30	0.28	0.29	0.29	0.30	0.26	0.25	0.26
	(98, 19)	d_VPH_PISOTI	[1 2 3 4 6]	0.66	0.51	0.45	0.36	0.33	0.30	0.30	0.27	0.31	0.27	0.29	0.28	0.27	0.26
	(98, 18)	d_PNACOE	[1 2 3 6 5]	0.68	0.52	0.49	0.36	0.38	0.33	0.31	0.30	0.32	0.33	0.31	0.29	0.26	0.28
	(98, 17)	d_VPH_SNBIEN	[1 2 3 6 5]	0.68	0.52	0.49	0.35	0.38	0.34	0.31	0.31	0.31	0.33	0.31	0.30	0.26	0.28
	(98, 16)	t_p8a14an_2010	[1 2 3 6 5]	0.68	0.53	0.51	0.37	0.39	0.35	0.35	0.32	0.34	0.34	0.30	0.28	0.29	0.28
	(98, 15)	d_P15YM_AN	[1 2 3 6 5]	0.68	0.54	0.51	0.38	0.39	0.37	0.33	0.29	0.33	0.33	0.30	0.27	0.29	0.27
	(98, 14)	d_P15YM_SE	[1 2 3 6 5]	0.68	0.56	0.50	0.38	0.39	0.36	0.32	0.32	0.32	0.34	0.30	0.27	0.29	0.27
	(98, 13)	d_PDESOCUP	[1 3 2 6 5]	0.68	0.58	0.54	0.42	0.41	0.33	0.30	0.30	0.31	0.32	0.29	0.30	0.28	0.27
	(98, 12)	t_psinder_2010	[1 3 2 6 4]	0.69	0.60	0.47	0.42	0.41	0.37	0.35	0.36	0.32	0.33	0.32	0.32	0.29	0.31
	(98, 11)	d_PSINDER	[1 3 2 6 4]	0.66	0.60	0.42	0.38	0.40	0.33	0.34	0.35	0.35	0.38	0.35	0.32	0.29	0.28
	(98, 10)	d_PROM_HNV	[1 3 2 6 4]	0.67	0.62	0.46	0.36	0.39	0.38	0.38	0.35	0.38	0.40	0.35	0.36	0.34	0.34
	(98, 9)	t_pnacoe_2010	[1 3 2 6 4]	0.66	0.59	0.53	0.43	0.43	0.47	0.43	0.40	0.39	0.43	0.42	0.40	0.40	0.39
	(98, 8)	t_pdesocup_2010	[3 1 2 5 4]	0.69	0.61	0.55	0.58	0.57	0.49	0.48	0.47	0.37	0.39	0.36	0.37	0.33	0.35
	(98, 7)	t_vph_snbien_2010	[3 1 2 5 4]	0.69	0.61	0.57	0.60	0.57	0.50	0.44	0.45	0.39	0.39	0.36	0.35	0.33	0.33
	(98, 6)	t_vph_pisot_2010	[5 3 2 1 4]	0.68	0.62	0.59	0.62	0.56	0.54	0.47	0.44	0.46	0.39	0.40	0.39	0.40	0.40
	(98, 5)	d_VPH_C_SERV	[5 2 3 1 4]	0.68	0.60	0.58	0.62	0.53	0.50	0.52	0.46	0.46	0.42	0.43	0.39	0.40	0.40
	(98, 4)	graproes_2010	[4 5 2 3 10]	0.75	0.63	0.69	0.64	0.60	0.50	0.47	0.47	0.43	0.43	0.42	0.39	0.37	0.36
	(98, 3)	prom_hnv_2010	[4 5 3 2 9]	0.71	0.68	0.67	0.61	0.54	0.50	0.49	0.46	0.44	0.47	0.44	0.45	0.44	0.41
	(98, 2)	t_p15ym_an_2010	[4 3 8 5 9]	0.67	0.67	0.67	0.50	0.53	0.46	0.47	0.47	0.49	0.47	0.46	0.45	0.48	0.46
	(98, 1)	t_p15ym_se_2010	[10 11 8 9 13]	0.65	0.62	0.59	0.61	0.60	0.60	0.64	0.62	0.65	0.65	0.63	0.61	0.59	0.58

Figura 4.53: Resultados al eliminar las variables de menor relevancia. Base CPV 98 localidades Nuevo León estandarizada.

Tomando como base las 13 variables más relevantes, se obtiene el grafo de similitudes y su correspondiente *embedding* espectral. Se realizan pruebas con distinto número de clústeres para garantizar buena interpretabilidad y finalmente se elige como el valor óptimo $k = 3$ clústeres.

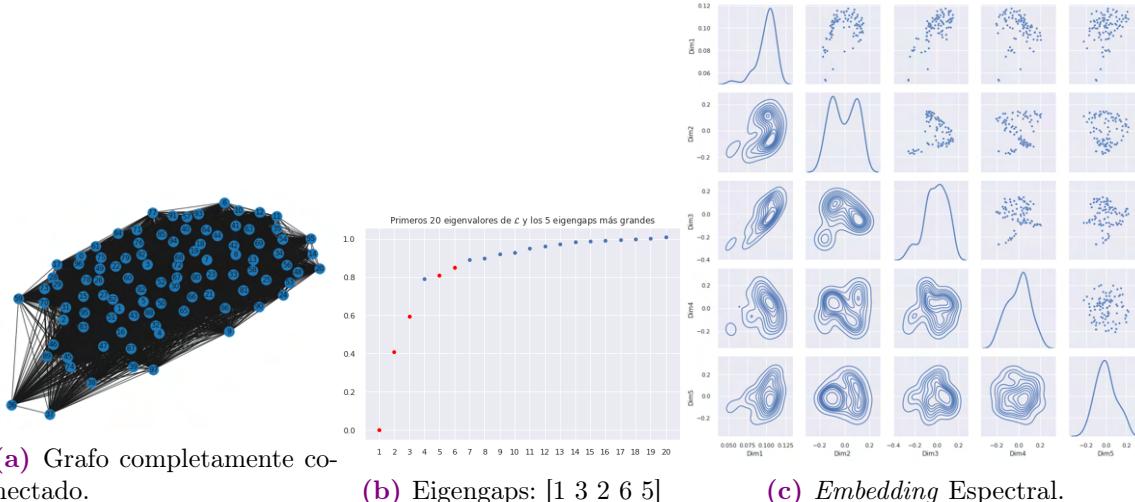


Figura 4.54: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base de 13 variables CPV, 98 localidades Nuevo León estandarizada.

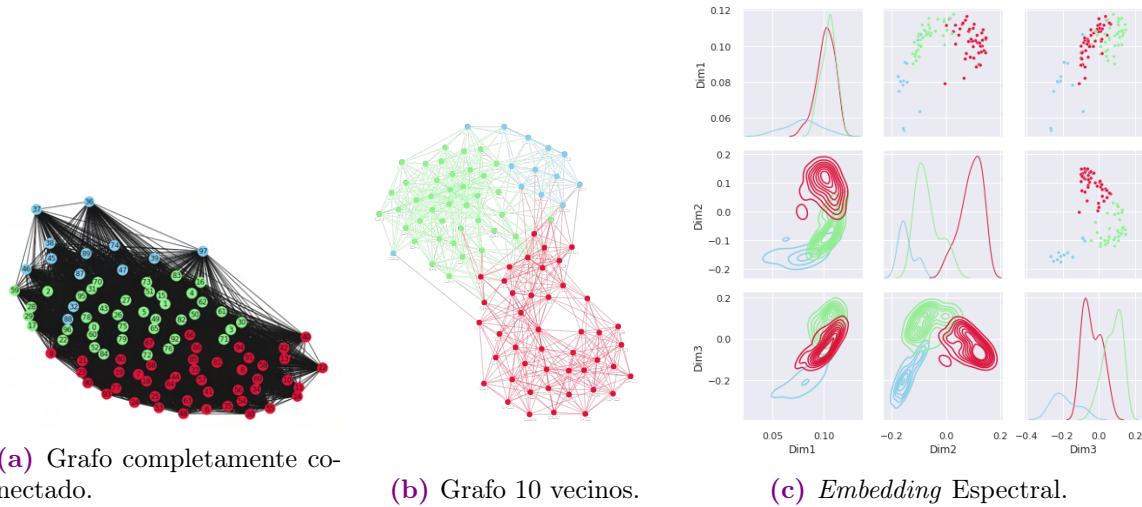


Figura 4.55: Resultados *clustering* espectral $k = 3$. Base 13 variables y 98 localidades Nuevo León estandarizada.

Eigensearch o Búsqueda iterativa del Eigengap. Clústeres DENUE

Después de haber elegido como versión final la agrupación con $k = 3$ y 13 variables del CPV, verificamos que en cada clúster ya no haya grupos por modelar. En la Figura 4.56 se observa que el primer gap de todos los clústeres es considerablemente mayor al resto de gaps, lo cual indica que ya no hay grupos dentro de los clústeres finales.

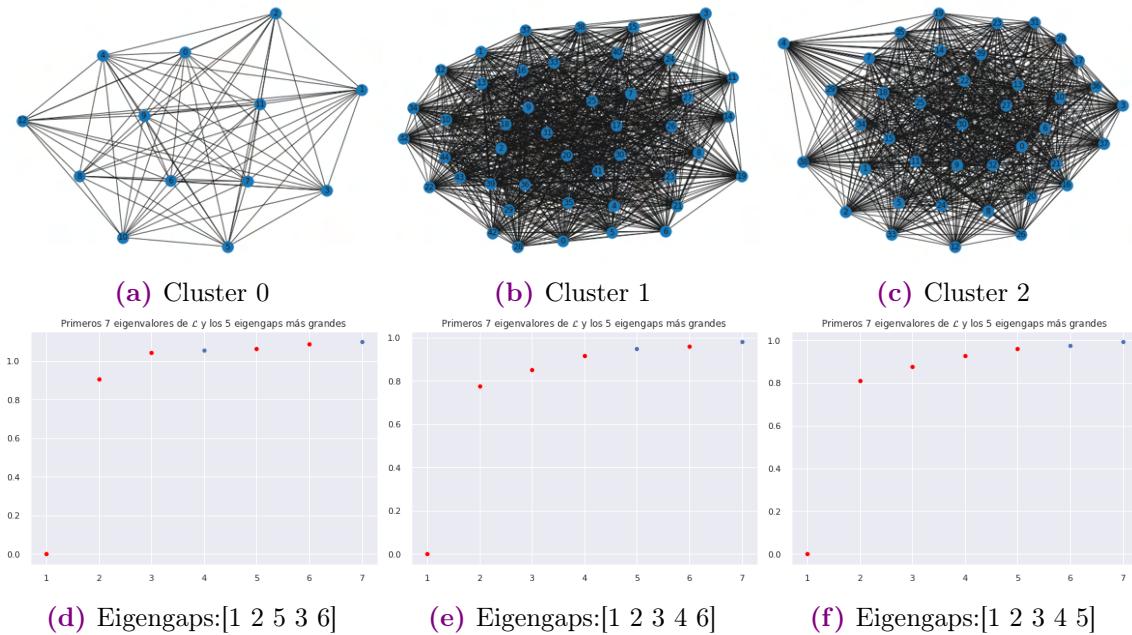


Figura 4.56: Eigensearch.

Interpretación Clústeres CPV

Los clústeres finales, considerando el conjunto de variables del CPV, quedaron conformados de la siguiente manera:

Cluster	Localidades	%
Rojo	45	46 %
Verde	40	41 %
Azul	13	13 %
Total	98	100 %

Tabla 4.9: Clústeres CPV Nuevo León.

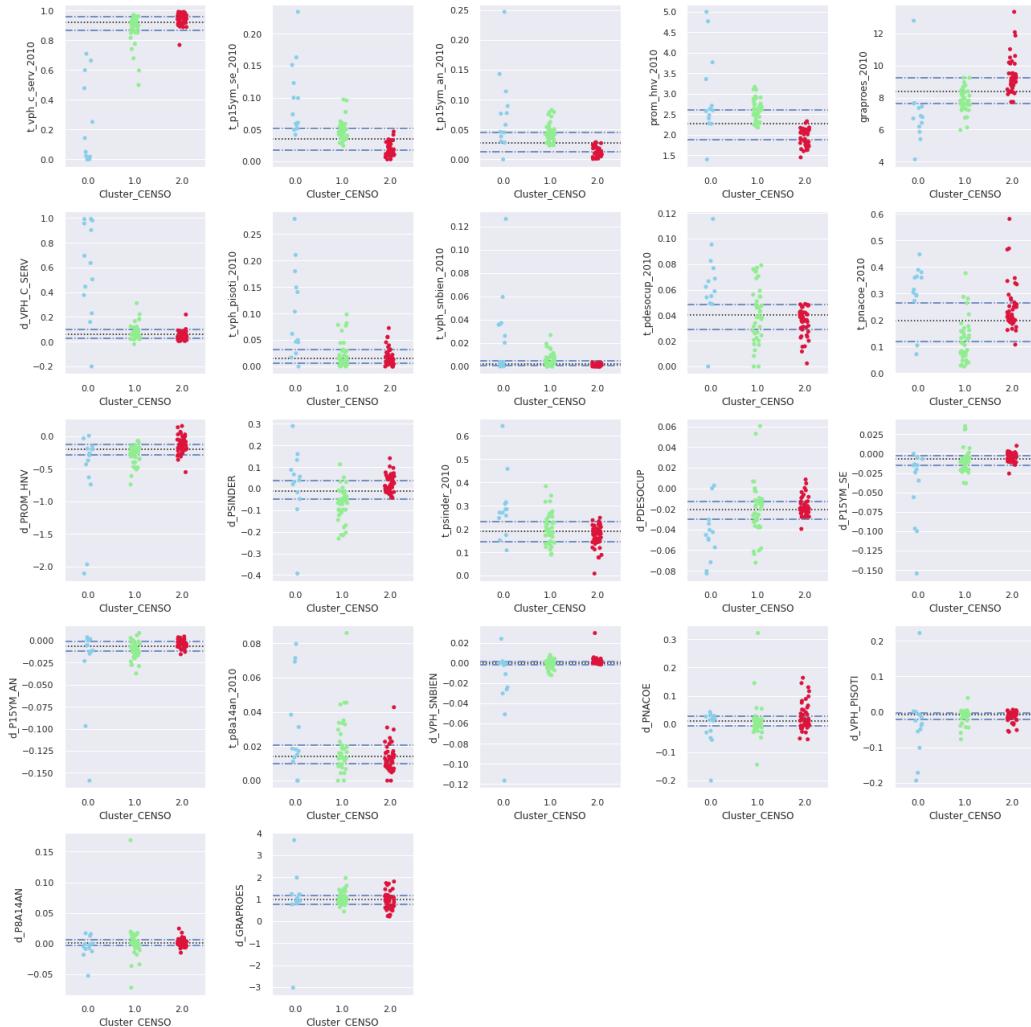


Figura 4.57: Dispersion clústeres CPV y cuantiles .25, .50 y .75 poblacionales.

En la Figura 4.58 se muestra la correlación lineal entre las variables y las dimensiones del *embedding*.

La dimensión 1 asocia valores negativos a localidades que en 2010 presentaban tasas altas de personas sin escolaridad, analfabetas, con viviendas de piso de tierra o viviendas sin bienes.

La dimensión 2 esta fuertemente relacionada positivamente a localidades con grados de escolaridad altos y cobertura amplia de viviendas que cuentan con servicios básicos. Por el contrario, valores negativos se asocian a analfabetismo y personas sin escolaridad acompañado de un promedio de hijos alto.

La dimensión 3 asocia valores positivos con localidades que en 2010 contaban con cobertura alta de servicios básicos y que continuaron mejorado en este aspecto en los últimos años. También se asocia positivamente a localidades donde se observa una disminución en el porcentaje de población sin afiliación a servicios de salud y que en 2010 presenta una tasa baja de población nacida en otra entidad.

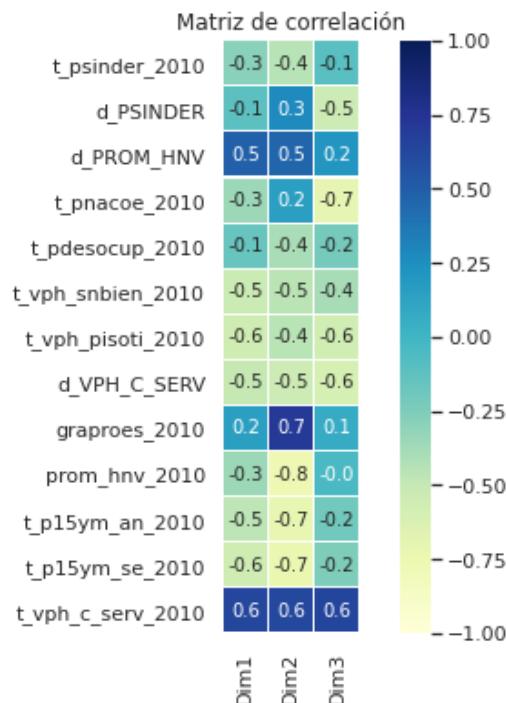


Figura 4.58: Matriz de correlación variables originales vs. *embedding*.



Figura 4.59: Dim1 vs. Dim2

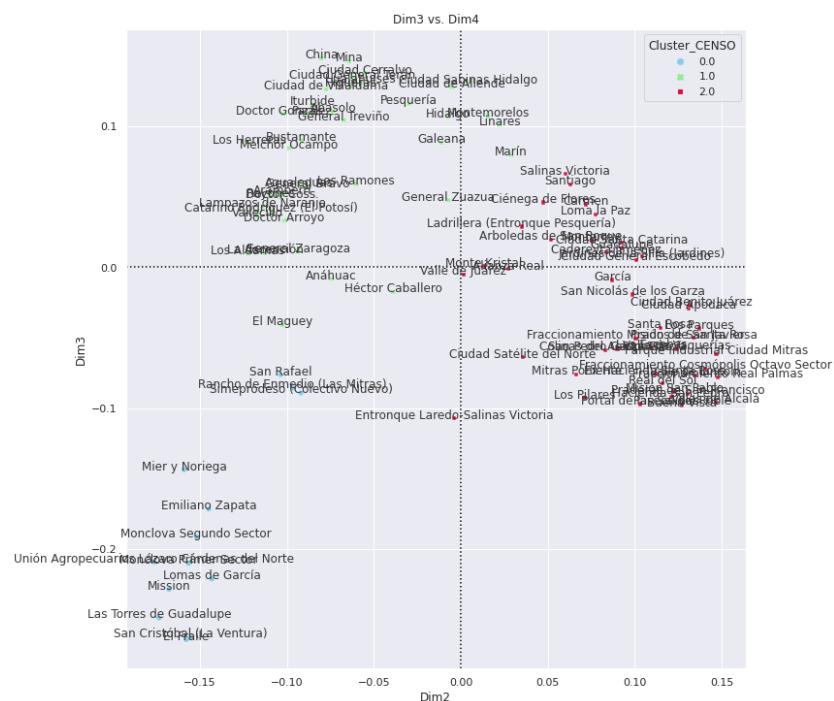


Figura 4.60: Dim2 vs. Dim3

Clúster Rojo:

- Este grupo presenta los mejores indicadores de bienestar social, en sus localidades, en promedio, el 95 % de las viviendas ya contaban con todos los servicios básicos desde 2010 y cuentan con los niveles más bajos de analfabetismo y personas sin escolaridad.
- Sus localidades cuentan con los grados de escolaridad más altos en comparación con el resto de localidades. Destacan Mitrás Poniente, San Pedro Garza García, Ex-Hacienda Santa Rosa, San Nicolás de los Garza, Monterrey, Ciudad Apodaca, entre otros.
- El promedio de hijos vivos es de los más bajos, al igual que la tasa de desocupación y tasa de población sin afiliación a servicios de salud.
- En los últimos años ha incrementado el porcentaje de población nacida en otra entidad y que habita en las localidades de este grupo.

Clúster Verde:

- Este grupo presenta indicadores de bienestar en una rango de medio a malo con respecto al resto de localidades. Presentan mayores niveles de población sin escolaridad y tasa de analfabetismo en comparación con el clúster rojo.
- En los últimos años la población sin afiliación a servicios de salud ha disminuido, en 2020 presentan las menores tasas de población sin afiliación.
- El grado de escolaridad se encuentra dentro de los niveles medios respecto al resto de localidades
- El promedio de hijos vivos es de los más altos junto con el clúster azul.
- Es el grupo con menor porcentaje de habitantes nacidos en otra entidad.

Clúster Azul:

- En este grupo se encuentran las localidades que aún en 2020 no cuentan con un porcentaje alto de viviendas con servicios básicos, algunas de ellas son El Fraile, Unión Agropecuarios Lázaro Cárdenas del Norte, Las Torres de Guadalupe, Mier y Noriega, entre otras.
- Presentan los niveles más altos de analfabetismo y población sin escolaridad.
- Tienen los grados de escolaridad más bajos respecto al resto de localidades y no ha mejorado en los últimos años.
- El promedio de hijos vivos es de los más altos junto con el clúster verde.
- Tiene los porcentajes más altos de población sin afiliación a servicios de salud.
- Sus localidades tienen un porcentaje alto de habitantes nacidos en otras entidades.

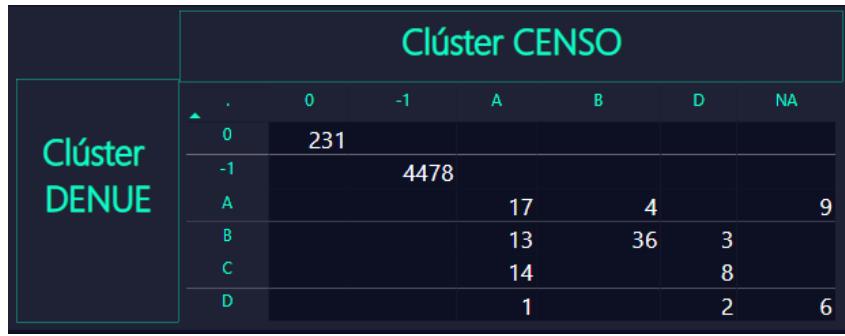
Con base en lo anterior se propone el siguiente *ranking* por nivel de bienestar social, donde 4 estrellas es el máximo nivel y una estrella el nivel mínimo:

Cluster CPV	Nivel de Bienestar
Rojo	★★★
Verde	★★
Azul	*

Tabla 4.10: Ranking Clústeres CPV Nuevo León.

Una vez obtenido el *ranking* tanto del conjunto de variables del DENUE como del CPV es posible identificar qué localidades presentan un escenario favorecedor para los micro negocios y además cuentan con los mejores niveles de bienestar en la entidad. Dichas localidades se encuentran rankeadas de acuerdo a la suma total de estrellas, por lo que el mejor clúster será la categoría A-A de 8 estrellas en total y posteriormente la categoría A-B o B-A con 7 estrellas, y así sucesivamente.

En el estado de Nuevo León son 17 las localidades con categoría A-A. Estas localidades se ubican en municipios aledaños a la capital del estado y parece expandirse

**Figura 4.61:** Localidades Nuevo León por categoría.

hacia el norte del estado en municipios como Salinas Victoria, El Carmen, General Zuazua y Apodaca.

**Figura 4.62:** Localidades con mayor potencial de crecimiento micro y bienestar social Nuevo León.

En general, en la mayoría de localidades predomina el sector de *Comercio al por menor*, en específico tiendas de abarrotes o misceláneas, y el sector de *Servicios de preparación de alimentos y bebidas*. Sin embargo, hay ciertos sectores que diferencian a las localidades. Por ejemplo, en Ciudad de Apodaca, Santa Catarina y San Nicolás

de los Garza predominan las UE en forma de papelerías, salones de belleza y servicios de reparación mecánica de automóviles y camiones.

En Las Escobas y Misión San Pablo, además del comercio al por menor, la industria manufacturera alimentaria tiene presencia. En Monterrey, se observa mayor diversidad de sectores pero la industria manufacturera para la fabricación de productos de herrería y comercio al por mayor de materias primas agropecuarias tiene fuerte presencia.

En San Pedro Garza García el comercio al por menor esta más enfocado a artículos de uso personal, de cuidado de la salud o de esparcimiento. En Santiago, destaca la decoración de interiores, comercio de muebles, artículos para el esparcimiento y de uso personal, accesorios de vestir y de calzado, industria manufacturera de muebles y de elaboración de pan.

En Valle de Vaquerías se tiene presencia de la industria manufacturera para elaboración de tortillas y productos de herrería.

Cabe destacar que, a diferencia de las localidades del estado de Nayarit, se observa que en Nuevo León ciertas localidades son mucho más grandes en extensión territorial como es el caso de la capital, para estos casos es mucho más conveniente realizar el análisis a nivel AGEB o manzana, ya que dentro de esta localidad existen diferencias muy notables a muy cortas distancias. Sin embargo, el alcance de este trabajo es visualizar a nivel más general para poder identificar zonas de potencial crecimiento que no necesariamente se encuentren en las zonas de mayor centralización del estado.

Por último, tomando como base las 98 localidades que pertenecen a alguna de las categorías (A,B,C,D) se obtuvo el siguiente mapa a nivel municipio iluminado de acuerdo al promedio de puntaje de las localidades que lo constituyen. En primer lugar se ubican Monterrey, San Nicolás de los Garza, San Pedro Garza García, Santa Catarina y Santiago; y en los últimos lugares se encuentran Mier y Noriega, General Escobedo, García y Galeana.

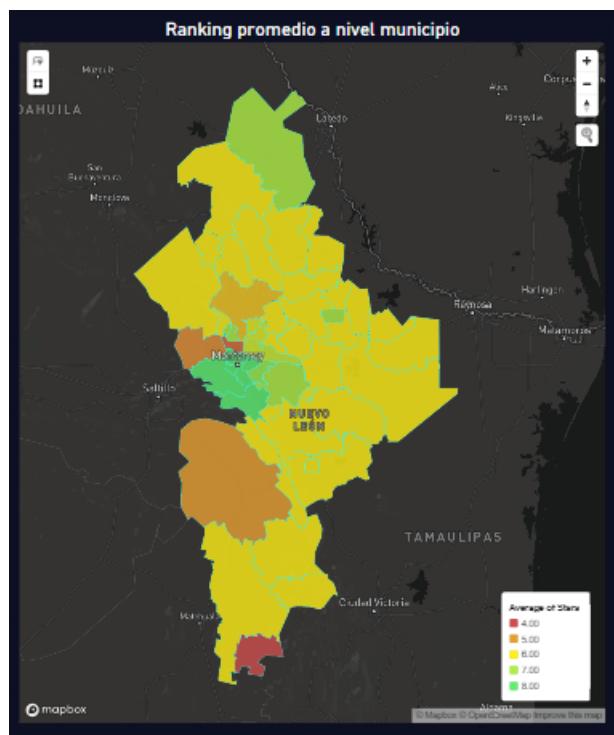


Figura 4.63: Ranking por municipio Nuevo León.

4.3. Yucatán

4.3.1. Análisis exploratorio

El estado de Yucatán esta constituido por 106 municipios y 2434 localidades con más de 1 o 2 viviendas, el 73.2 % son rurales puntuales, 21.3 % rurales y 5.5 % urbanas. Gran parte sus localidades (91.3 %) no tienen ningún registro de UE Micro en el DENUE, por lo tanto se colocan dentro del *Cluster -1* de nula actividad económica del sector Micro.

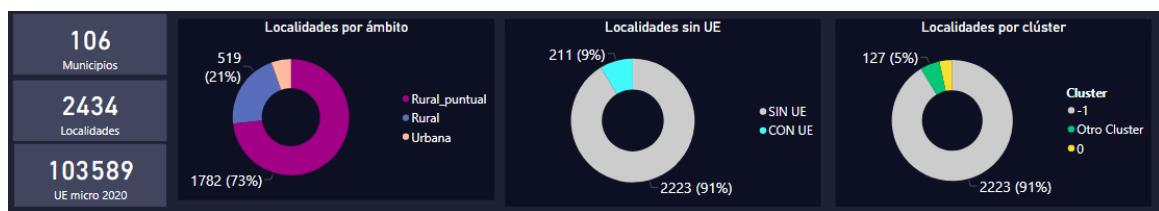


Figura 4.64: Información general Yucatán

Las 211 localidades que si cuentan con UE se localizan en la zona Noroeste del estado (Figura 4.65).



Figura 4.65: Localidades con Unidades Económicas Yucatán

Cabe mencionar que de estas 211 localidades, 84 registraron 5 o menos UE activas en 2020, por lo que estas localidades son agrupadas en el *Cluster 0* de baja actividad

económica micro.

A continuación se muestran las localidades que quedaron agrupadas en el *Cluster -1* (gris), *Cluster 0* (amarillo) y Otros (verde).

Cluster	Descripción	Localidades	Urbanas	Rurales	Rurales puntuales
-1	Nula actividad micro	2,223	4	450	1,769
0	Baja actividad micro	84	3	68	13
Otros	Loc. por clusterizar	127	126	1	0

Tabla 4.11: Distribución localidades Yucatán. Cluster -1, 0 y Otros.

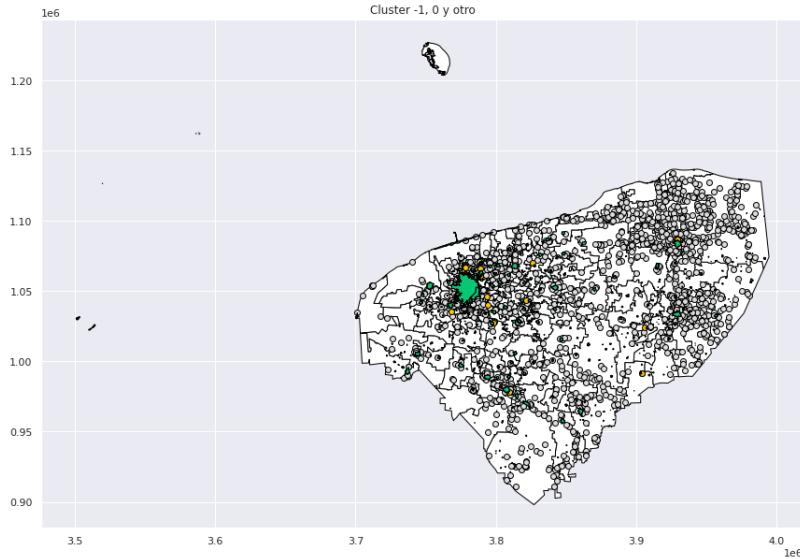


Figura 4.66: Distribución localidades Yucatán. Cluster -1, 0 y Otros.

4.3.2. Selección de variables mediante algoritmo SPEC.

Considerando la base de datos de las 127 localidades a clusterizar, se aplica el algoritmo SPEC para identificar las variables que tienen mayor influencia en la formación de los clústeres. Los parámetros utilizados y función de rankeo son los mismos que se utilizaron para los datos de Nayarit y Nuevo León, por lo tanto valores pequeños en el score $\varphi_2(F_i)$ indican que la variable F_i se alinea estrechamente con los eigen-vectores no triviales de los eigenvalores pequeños y por lo tanto la i -ésima variable provee una buena separabilidad de las observaciones.

El *ranking* de variables se realiza para el conjunto de variables del DENUE y del CPV por separado ya que de esta manera se obtuvieron mejores resultados. Los resultados para cada conjunto se muestran a continuación.

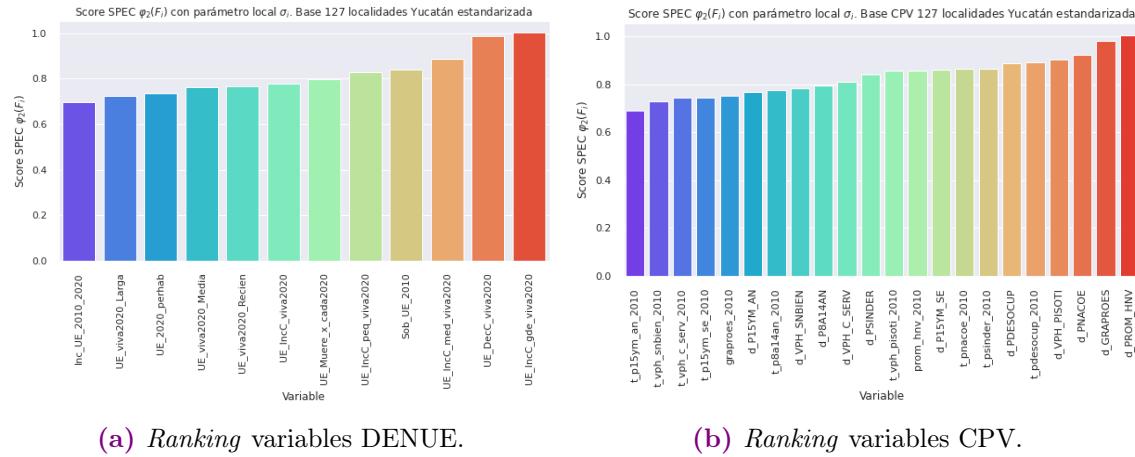


Figura 4.67: Resultados algoritmo SPEC con parámetro local σ_i . Base DENUE y base CPV Yucatán estandarizada. *Ranking* de mayor a menor relevancia.

En la figura 4.67a se observa que las variables que más influyen en la clusterización son las referentes al número de UE por habitante y la antigüedad de las UE. Por otra parte, las variables de menor relevancia son las variables que miden la proporción de negocios micro que incrementaron de personal pasando a ser medianas y grandes empresas y aquellas que decrementaron en número de personal. La poca relevancia de estas variables se debe a que son muy pocas las localidades que presentan este tipo de casos de incremento o decrecimiento de personal, por lo tanto no tienen un poder discriminante en las observaciones.

En el conjunto de las variables del CPV, se observa que las primeras 10 variables son las mayor relevancia, ya que a partir de éstas el score $\varphi_2(F_i)$ se mantiene casi constante (Figura 4.67b).

4.3.3. *Clustering* Espectral

Para encontrar la estructura final de los clústeres se aplica la misma metodología que se utilizó en los datos de los dos estados anteriores. Se obtienen clústeres por separado para las variables del DENUE y del CPV, para cada caso se obtienen las

agrupaciones considerando todas las variables y en caso de no encontrar un buen ajuste e interpretabilidad se eliminan las variables de menor relevancia hasta encontrar la mejor opción de *clustering*.

A continuación se describen los resultados obtenidos.

Grafo completamente conectado, 12 variables DENUE.

Al considerar todas las variables del DENUE, el grafo de similitudes no revela la existencia de grupos y esto se ve reflejado en los eigengaps, la brecha más grande la observamos en el primer eigenvalor lo cual indica que el número de clústeres sea 1. Por otra parte, en el *embedding* espectral se logra identificar a lo más 2 clústeres, ya que a partir de la tercera dimensión la dispersión de puntos forma una nube homogénea sin tendencias visibles.

Se realiza el proceso de *clustering* considerando distinto número de grupos, sin embargo, los grupos formados tienen mucha distorsión o en el caso de $k = 2$ se tienen clústeres muy generalizados que no revelan información interesante.

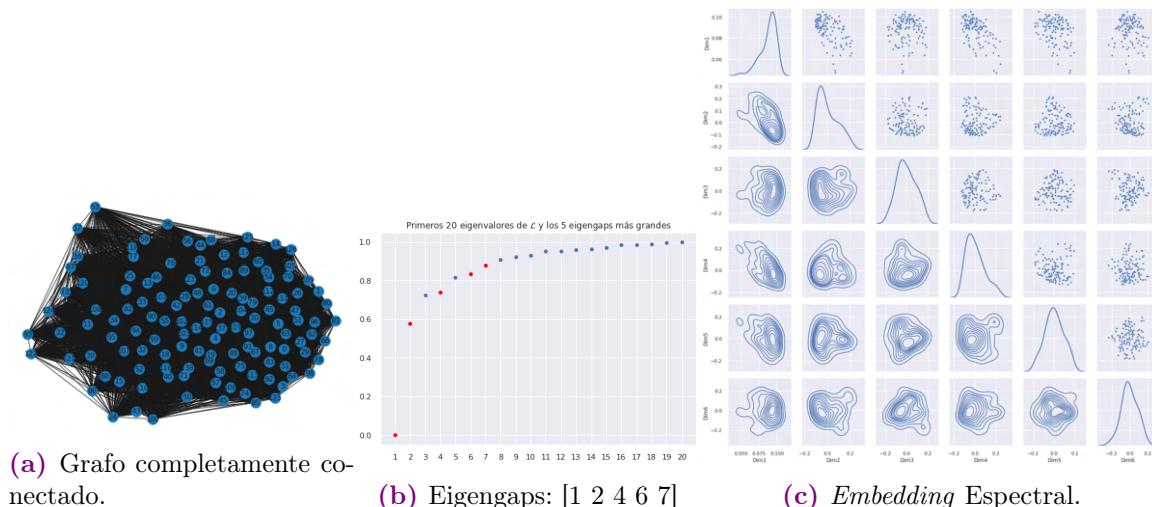


Figura 4.68: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base de 12 variables DENUE, 127 localidades Yucatán estandarizada.

Por lo anterior, se decide verificar si al eliminar las variables de menor relevancia se obtiene un grafo más claro y en consecuencia clústeres más definidos.

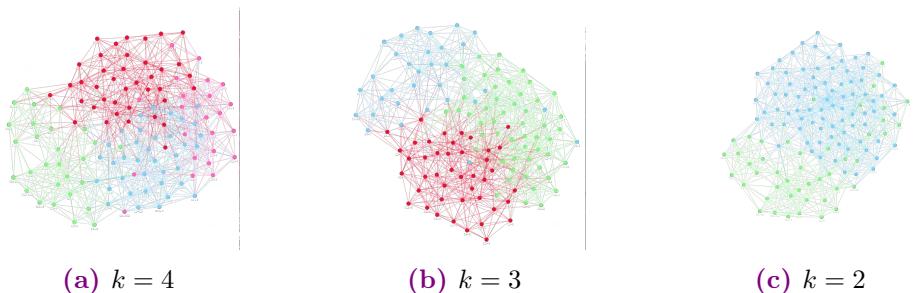


Figura 4.69: clustering $k = 4$, $k = 3$ y $k = 2$. Base de 12 variables DENUE, 127 localidades Yucatán estandarizada.

**Grafo completamente conectado, eliminando variables de menor relevancia
DENUE.**

Se realiza el ejercicio de eliminar de una en una las variables de menor relevancia.

En la Figura 4.70 se observa una mejora en el estadístico Silhouette y el heurístico eigengap al eliminar las 7 variables de menor relevancia.

Además, el segundo gap más grande pasa de 2 a 3 indicando que posiblemente haya 3 clústeres en el grafo y el criterio Silhouette mejora significativamente al elegir en $k = 3$, $k = 4$ o $k = 5$ clústeres.

Score	Dimensiones	V. Eliminada	Eigengap Top5	Número de Clusters												
				2	3	4	5	6	7	8	9	10	11	12	13	
Silhouette	(127, 12)	Ninguna	[1 2 4 6 7]	0.66	0.40	0.33	0.33	0.32	0.31	0.30	0.30	0.28	0.27	0.26	0.26	0.24
	(127, 11)	UE_IncG_gde_viva202	[1 2 4 6 7]	0.66	0.44	0.33	0.32	0.32	0.31	0.29	0.29	0.27	0.27	0.25	0.25	0.22
	(127, 10)	UE_DecC_viva2020	[1 2 4 6 7]	0.67	0.42	0.33	0.31	0.31	0.30	0.29	0.29	0.28	0.27	0.26	0.25	0.22
	(127, 9)	UE_IncC_med_viva202	[1 2 4 6 7]	0.69	0.44	0.34	0.35	0.33	0.32	0.27	0.27	0.24	0.23	0.25	0.22	0.24
	(127, 8)	Sob_UUE_2010	[1 2 4 6 5]	0.71	0.47	0.35	0.37	0.35	0.32	0.30	0.28	0.25	0.24	0.24	0.23	0.22
	(127, 7)	UE_IncP_preq_viva202	[1 2 3 4 6]	0.72	0.47	0.37	0.38	0.35	0.29	0.27	0.26	0.26	0.27	0.27	0.25	0.24
	(127, 6)	UE_Muere_x_cada202	[1 2 3 4 6]	0.73	0.48	0.41	0.40	0.39	0.36	0.32	0.32	0.30	0.30	0.28	0.27	0.25
	(127, 5)	UE_IncP_viva2020	[1 3 5 2 4]	0.67	0.53	0.51	0.50	0.43	0.37	0.37	0.37	0.36	0.34	0.33	0.32	0.32
	(127, 4)	UE_viva2020_Reden	[2 1 3 7 4]	0.76	0.55	0.53	0.48	0.43	0.39	0.37	0.35	0.32	0.33	0.33	0.30	0.29
	(127, 3)	UE_viva2020_Media	[6 1 2 4 8]	0.72	0.53	0.57	0.54	0.47	0.53	0.48	0.48	0.44	0.43	0.37	0.36	0.35
Ari	(127, 2)	UE_2020_preamba	[3 6 4 1 2]	0.65	0.59	0.51	0.44	0.43	0.45	0.50	0.48	0.44	0.44	0.44	0.44	0.44
	(127, 1)	UE_viva2020_Larga	[12 8 10 11 14]	0.72	0.60	0.64	0.69	0.65	0.65	0.65	0.66	0.64	0.68	0.66	0.65	0.62

Figura 4.70: Resultados al eliminar las variables de menor relevancia. Base DENUE 127 localidades Yucatán estandarizada.

Tomando como base las 5 variables más relevantes, se obtiene el grafo de similitudes y su correspondiente *embedding* espectral, en este último se observan 3 poblaciones, por lo que $k = 3$ se perfila a ser el número óptimo de clústeres. No obstante se realizan pruebas con diferente número de clústeres para garantizar buena interpretabilidad, finalmente se elige como el valor óptimo $k = 4$ clústeres ya que aunque se pierde algunas décimas en el criterio Silhouette, hay una diferenciación mayor en los grupos y por lo tanto mejora la interpretabilidad.

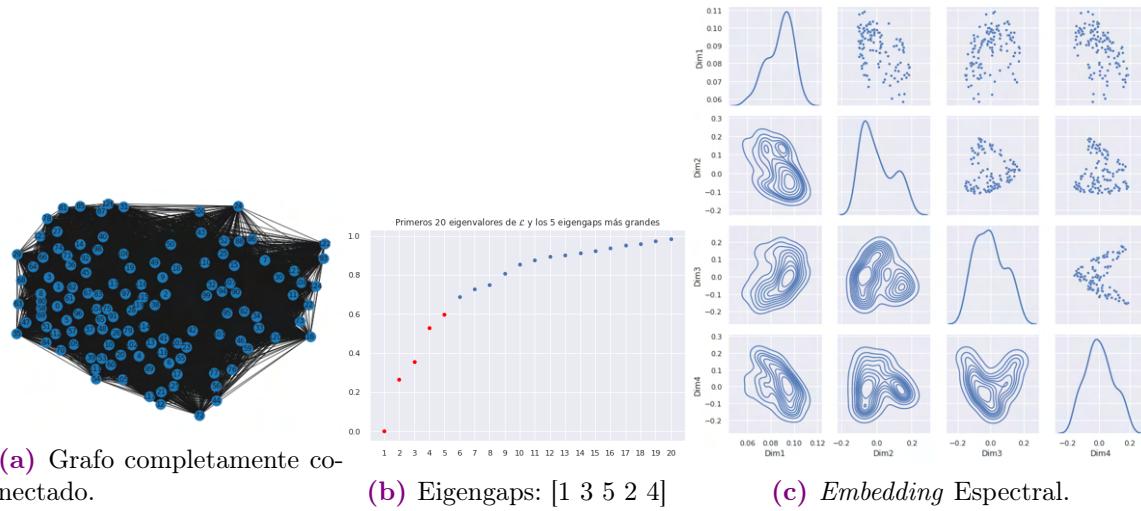


Figura 4.71: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base 5 variables DENUE, 127 localidades Yucatán estandarizada.

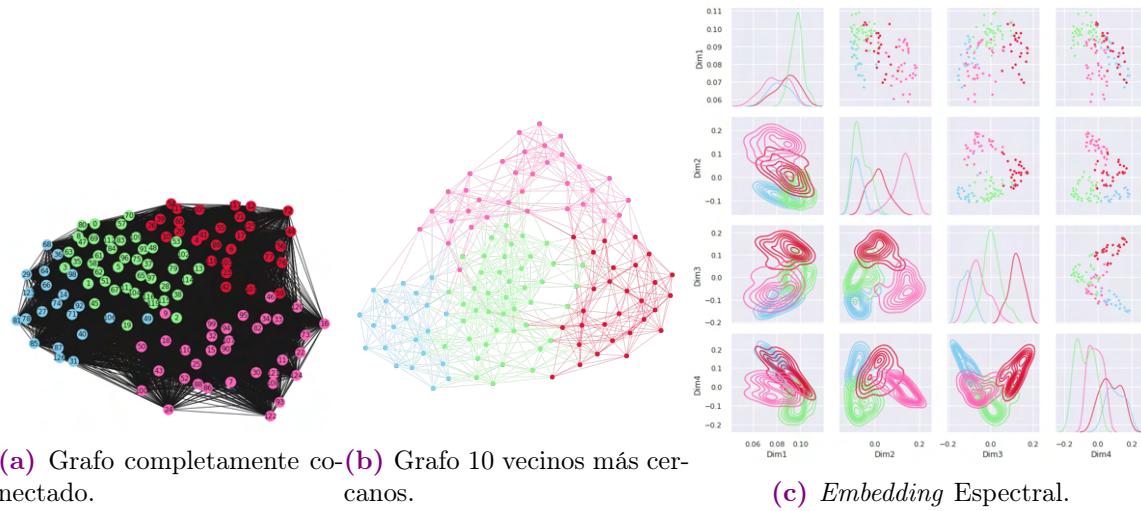


Figura 4.72: Resultados *clustering* espectral $k = 4$ con parámetro local σ_i . Base de 5 variables y 127 localidades Yucatán estandarizada.

Eigensearch o Búsqueda iterativa del Eigengap. Clústeres DENUE

Después de haber elegido como versión final la agrupación con $k = 4$ y 5 variables del DENUE, verificamos que en cada clúster ya no haya grupos por modelar.

En las siguientes gráficas vemos que el primer gap es considerablemente mayor al resto de gaps, lo cual indica que ya no hay más grupos dentro de los clústeres finales.

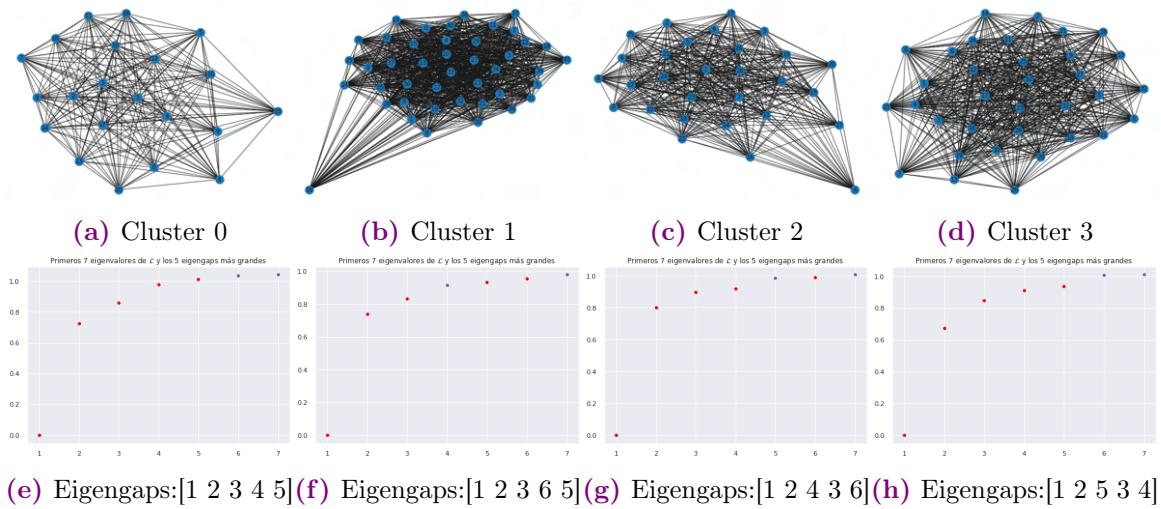


Figura 4.73: Eigensearch.

Interpretación Clústeres DENUE

Los clústeres finales, considerando el conjunto de variables del DENUE, quedaron conformados de la siguiente manera:

Cluster	Localidades	%
Verde	43	34 %
Rosa	34	27 %
Rojo	29	23 %
Azul	21	16 %
Total	127	100 %

Tabla 4.12: Clústeres DENUE Yucatán.

A continuación se realiza la interpretación de clústeres con base en la dispersión de las variables originales, sus cuatiles poblacionales y el *embedding* espectral.

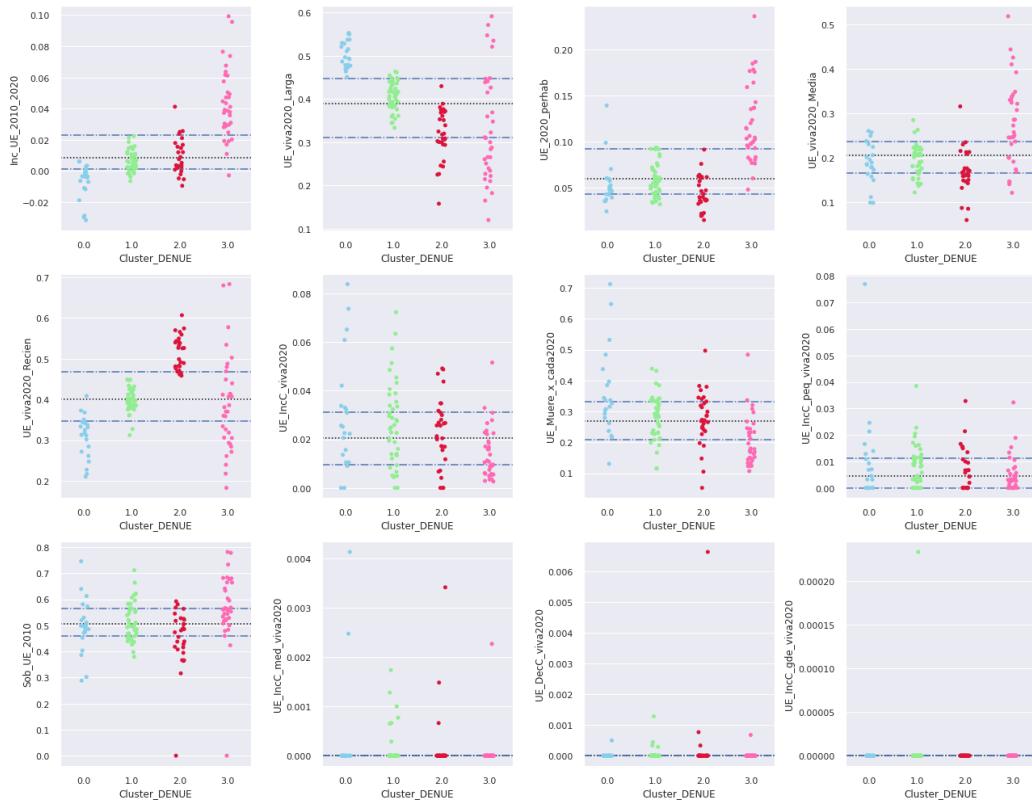


Figura 4.74: Dispersion clústeres DENUE y cuantiles .25, .50 y .75 poblacionales.

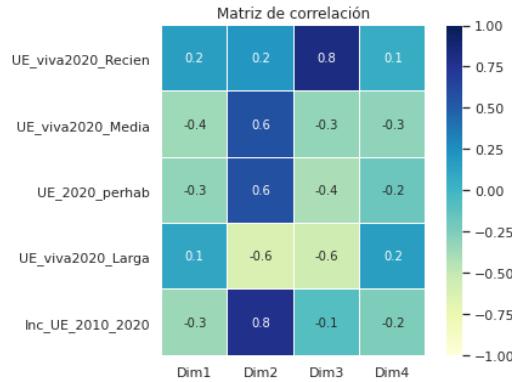


Figura 4.75: Matriz de correlación variables originales vs. *embedding*.

En la Figura 4.75 se muestra la correlación lineal entre las variables y las dimensiones del *embedding*.

La dimensión 1 presenta una débil correlación con todas las variables y esto se debe a que corresponde al eigenvector del primer eigenvalor, cuyo valor siempre es 0. Sin embargo, valores ligeramente más negativos se relacionan a localidades de antigüedad media que han presentado incrementos en cuanto al número de UE por habitante.

La dimensión 2 esta fuertemente asociada positivamente a localidades con antigüedades medias que han presentado incrementos en cuanto al número de UE por habitante.

La dimensión 3 se relaciona positivamente a localidades de antigüedad reciente que no presentan incrementos en el número de UE por habitante.

La dimensión 4 es muy similar a la 1, se relacionan negativamente a localidades de antigüedad media que han presentado incrementos en cuanto al número de UE por habitante. Esta última dimensión aporta muy poca información, sin embargo, se decide mantenerla ya que al tener 4 clústeres la distribución de los mismos en las variables originales es mucho más clara.

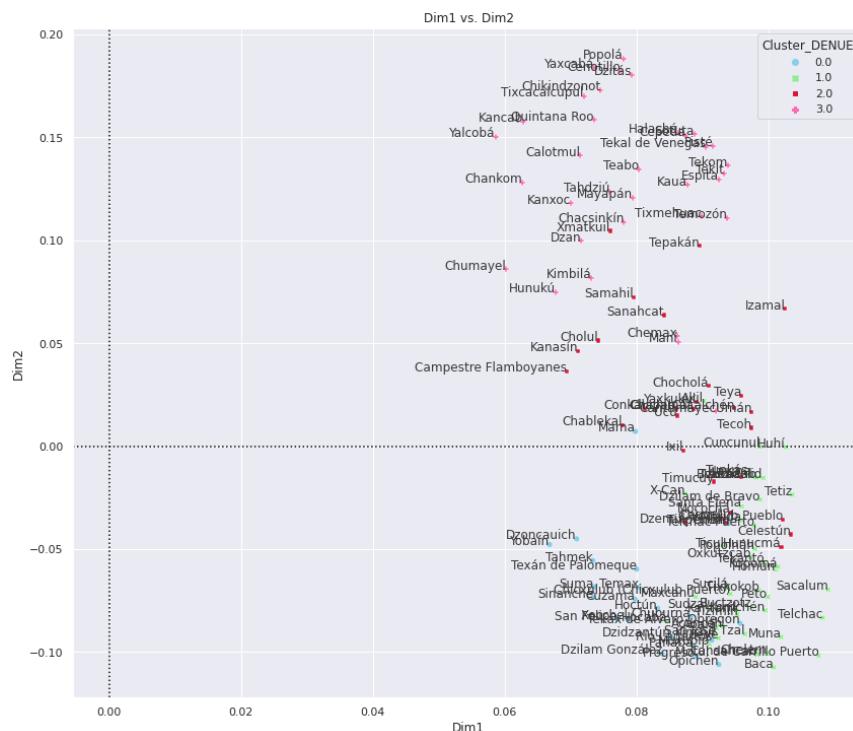


Figura 4.76: Dim1 vs. Dim2

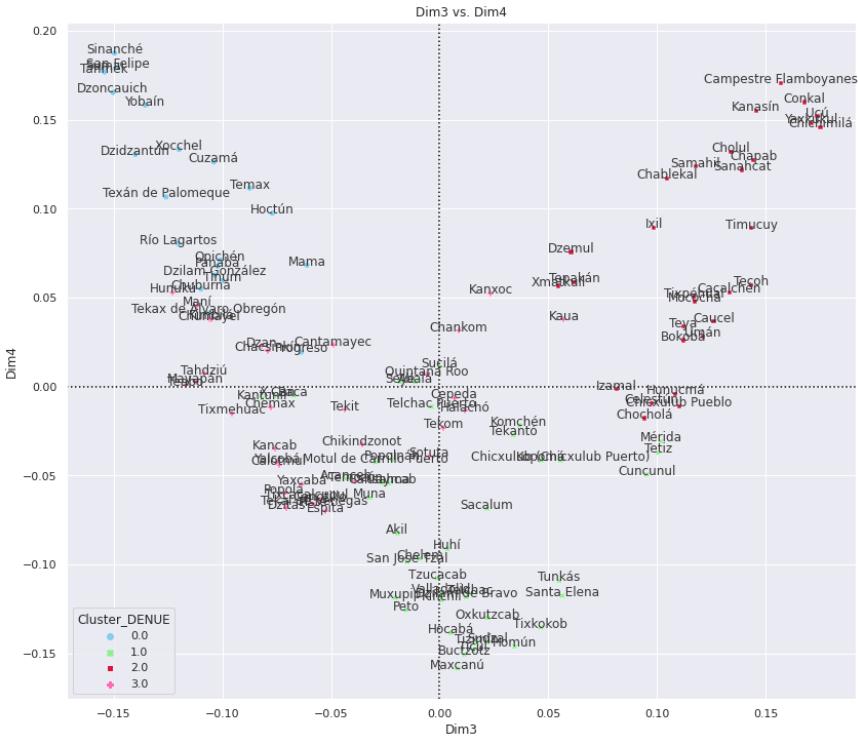


Figura 4.77: Dim3 vs. Dim4

Clúster Verde:

- Este clúster es el segundo con mayor número de UE de mayor antigüedad después del color azul, la antigüedad esta relacionada con los crecimientos de negocios micro a medianos ya que es el clúster con más crecimientos de este tipo.
- La proporción de UE de recién creación (creadas a partir de 2019) y de mediana creación (de 2014 a 2018) se encuentran dentro del nivel promedio.
- Los incrementos de UE por habitante se encuentran dentro de los niveles normales.

Por lo tanto, el clúster verde concentra a las localidades que tuvieron su mayor crecimiento antes de 2014 y que se han mantenido a un nivel promedio. En este clúster encontramos Mérida, Valladolid, Tizimím, Motul de Carrillo Puerto, Chelem, Sudzal, entre otros.

Clúster Rosa:

- Este clúster contiene a las localidades con los incrementos más altos de UE por habitante, por dichos incrementos, estas localidades se posicionan como las localidades con más UE por habitante en 2020.
- La composición por antigüedad de las UE es uniforme, cuenta con una proporción de alrededor de 30 % en los tres tipos de antigüedades.
- La sobrevivencia de las UE más antiguas (registradas en 2010) es la más alta en comparación al resto de localidades.
- El promedio de UE que mueren por cada unidad viva en 2020 es la más baja en comparación con los otros clústeres.

El clúster Rosa concentra a las localidades con mayor potencial de crecimiento, presenta un escenario favorecedor tanto para nuevos negocios como negocios con mayor antigüedad. Algunas de las localidades en este clúster son Chumayel, Calotmul, Sotuta, Kancab, etc.

Clúster Rojo:

- Este clúster contiene a las localidades con las UE más jóvenes ya que en promedio el 50 % de sus unidades se crearon en 2019.
- El número de UE por habitante se ha mantenido constante, lo cual sugiere que el número de negocios micro va en crecimiento a la par de la población.

Este clúster concentra a las localidades más jóvenes, sin embargo, no presenta un crecimiento fuera de los niveles normales. Entre las localidades se encuentra Tepakán, Izamal, Campestre Flamboyanes, Chocholá, Cholul, entre otras.

Clúster Azul:

- Las localidades de este clúster tienen el mayor número de UE creadas antes de 2014, por lo que, al igual que el clúster verde, se han dado crecimientos de negocios micro a pequeños y medianos.

- El número de UE por habitante ha decrecido en los últimos años y presenta los niveles más bajos de negocios creados a partir de 2019 respecto al resto de localidades.
- El promedio de UE que mueren por cada unidad viva en 2020 es la más alta en comparación con los otros clústeres.

Este clúster concentra a las localidades que tuvieron un crecimiento en años anteriores pero que en últimos años no ha presentado un escenario favorecedor e incluso podría encontrarse en declive.

Con base en lo anterior, se propone el siguiente *ranking* por nivel de potencial del sector micro, donde 4 estrellas es el máximo nivel y una estrella en nivel mínimo:

Cluster DENUE	Potencial sector Micro	Categoría
Rosa	★★★	A
Verde	★★	B
Rojo	★★	C
Azul	★	D

Tabla 4.13: *Ranking* Clústeres DENUE Yucatán.

Grafo completamente conectado, 22 variables CPV

Se obtiene el grafo de similitudes considerando las 22 variables del CPV, el resultado es una nube de observaciones homogénea en la que parece no haber clústeres. Así mismo, el eigengap más grande se encuentra en el primer eigenvalor indicando que el número de agrupaciones en el grafo es igual a 1.

En el *embedding* espectral, se puede apreciar que las primeras 3 dimensiones son las que discriminan mejor las observaciones ya que a partir de la cuarta dimensión se observa una nube de puntos sin formas claras.

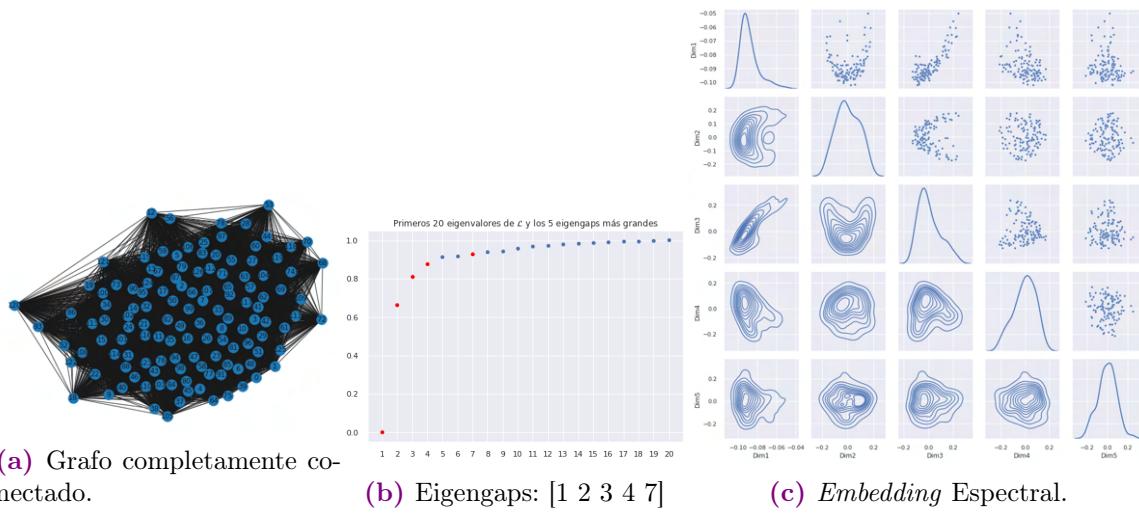


Figura 4.78: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base de 22 variables CPV, 127 localidades Yucatán estandarizada.

Bajo este escenario, se considera conveniente eliminar las variables de menor relevancia hasta el punto donde los clústeres comiencen a ser más evidentes.

Grafo completamente conectado, eliminando variables de menor relevancia CPV

Se realiza el ejercicio de eliminar de una en una las variables de menor relevancia de acuerdo al *ranking* del algoritmo SPEC. En la Figura 4.79 se observa una mejora importante en el estadístico Silhouette al eliminar las 9 variables de menor relevancia y elegir $k = 3$ clústeres. Si bien los eigengaps más grandes no modificaron su orden, se observó que el segundo y tercer eigengap incrementaron considerablemente con

respecto al grafo de similitudes construido a partir de las 22 variables, esto indicaría que el número óptimo de clústeres puede ser hasta $k = 3$ (Figura 4.80).

Por otra parte, cuando se eliminaron las últimas 12 variables menos relevantes se observó una mejora con $k = 4$ y $k = 5$, sin embargo, la ganancia en términos de interpretabilidad es mínima, por lo tanto, se decide elegir como versión final la base de 13 variables y $k = 3$ clústeres.

Valores	Dim	V. Eliminada	Eigengap Top5	Número de Clusters													
				2	3	4	5	6	7	8	9	10	11	12	13	14	15
Silhouette	(127, 22)	Ninguna	[1 2 3 4 7]	0.60	0.48	0.36	0.29	0.27	0.24	0.21	0.21	0.21	0.20	0.20	0.20	0.19	0.19
	(127, 21)	d_PROM_HNV	[1 2 3 4 9]	0.60	0.46	0.35	0.32	0.29	0.26	0.22	0.21	0.21	0.21	0.20	0.20	0.19	0.18
	(127, 20)	d_GRAPROES	[1 2 3 4 7]	0.61	0.48	0.36	0.33	0.30	0.26	0.24	0.21	0.21	0.21	0.19	0.20	0.19	0.18
	(127, 19)	d_PNAOCOE	[1 2 3 4 7]	0.60	0.47	0.36	0.33	0.28	0.26	0.23	0.22	0.22	0.20	0.21	0.19	0.18	0.18
	(127, 18)	d_VPH_PISOTI	[1 2 3 4 6]	0.60	0.49	0.37	0.33	0.28	0.25	0.22	0.24	0.21	0.21	0.20	0.18	0.19	0.17
	(127, 17)	t_pdescup_2010	[1 2 3 4 6]	0.60	0.50	0.38	0.34	0.31	0.27	0.23	0.23	0.22	0.20	0.20	0.19	0.19	0.17
	(127, 16)	d_PDESOCUP	[1 2 3 4 6]	0.60	0.50	0.39	0.35	0.31	0.28	0.26	0.25	0.24	0.21	0.21	0.20	0.18	0.17
	(127, 15)	t_psinder_2010	[1 2 3 5 4]	0.61	0.51	0.41	0.38	0.31	0.28	0.24	0.24	0.23	0.23	0.21	0.20	0.18	0.19
	(127, 14)	t_pnaocoe_2010	[1 2 3 5 7]	0.59	0.54	0.41	0.39	0.32	0.28	0.28	0.24	0.25	0.23	0.20	0.20	0.19	0.20
	(127, 13)	d_P15YM_SE	[1 2 3 6 4]	0.58	0.56	0.43	0.39	0.35	0.35	0.30	0.25	0.26	0.23	0.23	0.24	0.21	0.22
	(127, 12)	prom_hnv_2010	[1 2 3 6 5]	0.60	0.54	0.43	0.39	0.36	0.34	0.30	0.28	0.26	0.26	0.27	0.24	0.23	0.23
	(127, 11)	t_vph_pisot_2010	[1 2 3 6 5]	0.61	0.54	0.43	0.39	0.35	0.34	0.31	0.30	0.27	0.26	0.26	0.25	0.24	0.24
	(127, 10)	d_PSINDER	[1 2 3 5 6]	0.61	0.54	0.48	0.43	0.42	0.33	0.31	0.31	0.27	0.26	0.27	0.24	0.25	0.23
	(127, 9)	d_VPH_C_SERV	[1 2 3 4 6]	0.59	0.55	0.51	0.46	0.41	0.36	0.30	0.32	0.30	0.28	0.27	0.24	0.24	0.24
	(127, 8)	d_P8A14AN	[1 2 3 4 6]	0.60	0.56	0.50	0.44	0.45	0.37	0.33	0.32	0.29	0.28	0.27	0.26	0.23	0.24
	(127, 7)	d_VPH_SNBIEN	[1 3 2 4 6]	0.60	0.57	0.52	0.45	0.43	0.34	0.35	0.30	0.28	0.27	0.24	0.23	0.24	0.24
	(127, 6)	t_p8a14an_2010	[3 1 2 4 6]	0.62	0.58	0.55	0.47	0.44	0.36	0.34	0.33	0.32	0.28	0.26	0.26	0.25	0.24
	(127, 5)	d_P15YM_AN	[3 2 1 4 6]	0.62	0.59	0.55	0.46	0.46	0.40	0.39	0.36	0.35	0.34	0.32	0.32	0.30	0.29
	(127, 4)	graproes_2010	[3 2 1 5 6]	0.66	0.60	0.58	0.52	0.47	0.45	0.42	0.36	0.35	0.36	0.35	0.34	0.35	0.32
	(127, 3)	t_p15ym_se_2010	[5 3 1 2 6]	0.60	0.60	0.56	0.47	0.49	0.46	0.40	0.38	0.35	0.36	0.38	0.37	0.33	0.31
	(127, 2)	t_vph_c_serv_2010	[4 5 10 7 3]	0.65	0.56	0.61	0.58	0.57	0.52	0.51	0.50	0.55	0.54	0.53	0.50	0.47	0.48
	(127, 1)	t_vph_snbien_2010	[9 13 7 12 19]	0.69	0.71	0.58	0.61	0.67	0.71	0.71	0.71	0.64	0.64	0.65	0.65	0.64	0.60

Figura 4.79: Resultados al eliminar las variables de menor relevancia. Base CPV 127 localidades Yucatán estandarizada.

Tomando como base las 13 variables más relevantes, se obtiene el grafo de similitudes y su correspondiente *embedding* espectral. En este último se observa que a partir de la cuarta dimensión no se observan distintas poblaciones en las estimaciones de densidad, lo cual confirma que el número óptimo de clústeres sea $k = 3$.

El resultado del *clustering* se muestra en la Figura 4.81. Se observa que hay 2 poblaciones (rojo y verde) en los extremos del grafo que se diferencian perfectamente una de la otra, pero la tercera clase que se encuentra en medio de las dos anteriores compartirá similitudes tanto con el clúster rojo como con el verde.

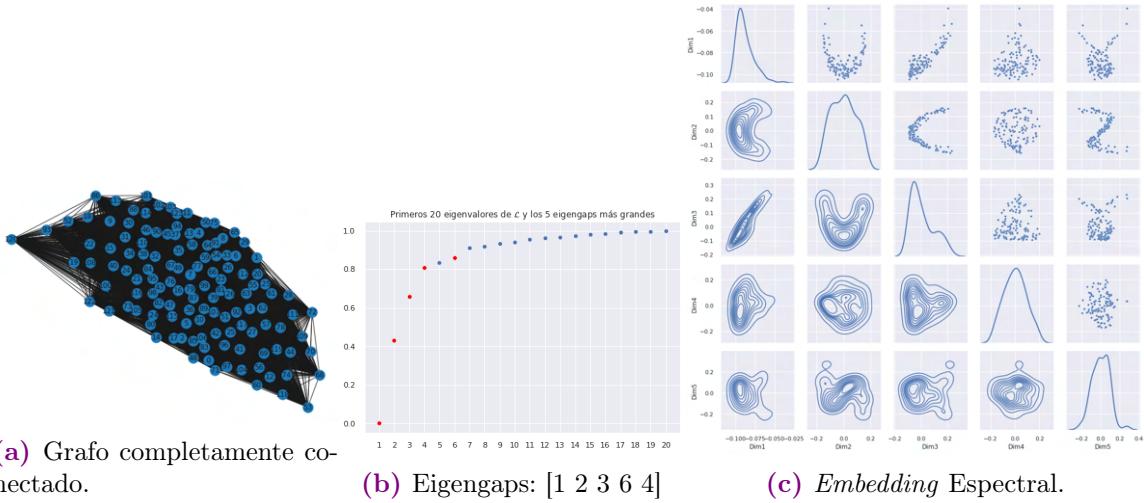


Figura 4.80: Grafo de similitudes, top 5 eigengaps y *embedding* espectral. Base de 13 variables CPV, 127 localidades Yucatán estandarizada.

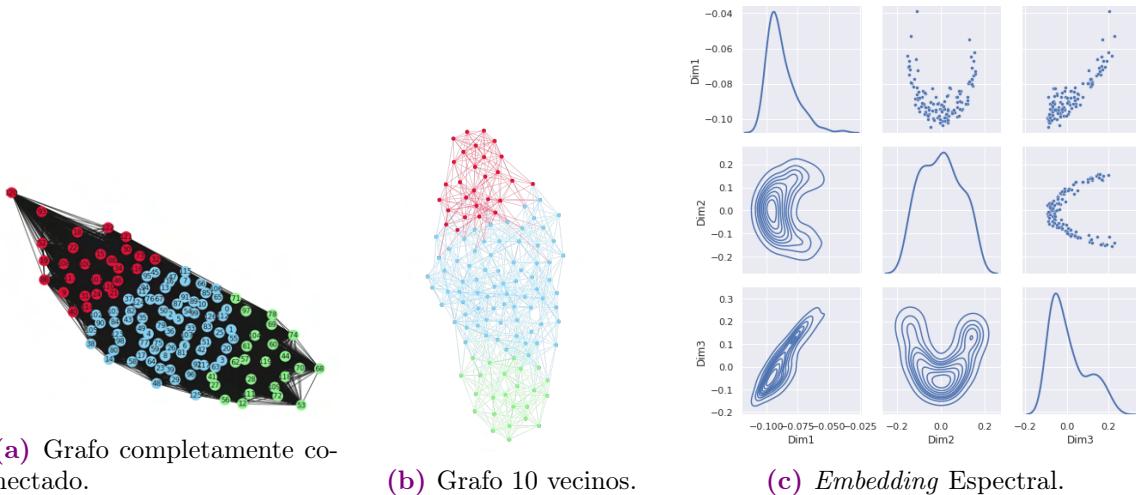
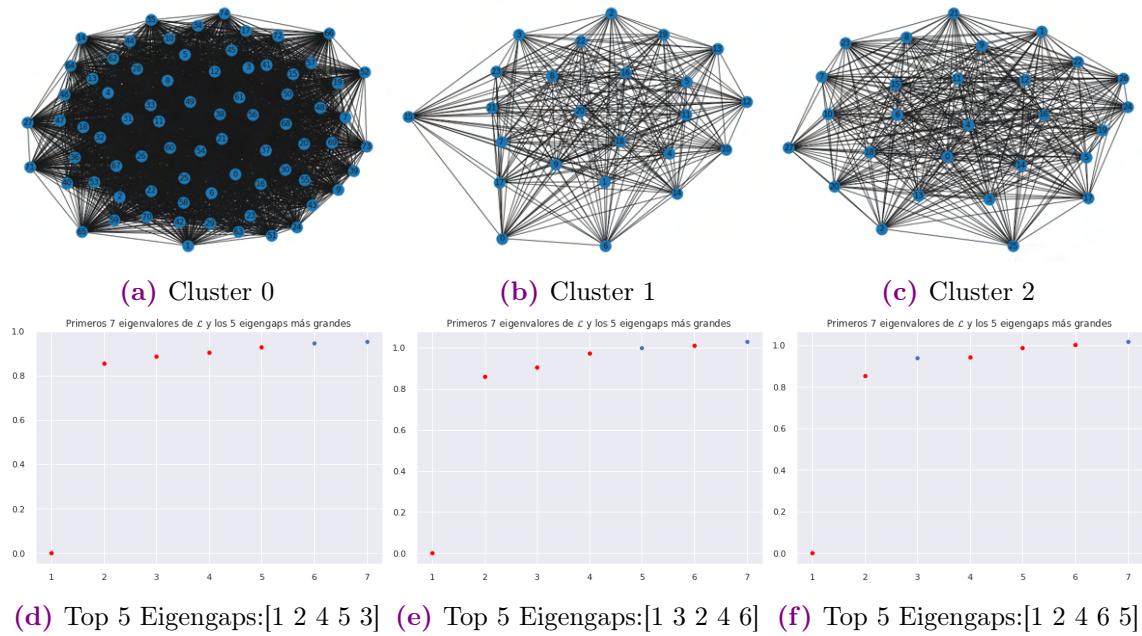


Figura 4.81: Resultados *clustering* espectral $k = 3$. Base 13 variables y 127 localidades Yucatán estandarizada.

Eigensearch o Búsqueda iterativa del Eigengap. Clústeres CPV.

Después de haber elegido como versión final la agrupación con $k = 3$ y 13 variables del CPV, verificamos que en cada clúster ya no haya grupos por modelar. En la Figura 4.82 se observa que el primer gap de todos los clústeres es considerablemente mayor al resto de gaps, lo cual indica que ya no hay grupos dentro de los clústeres finales.

**Figura 4.82:** Eigensearch.

Interpretación Clústeres CPV

Los clústeres finales, considerando el conjunto de variables del CPV, quedaron conformados de la siguiente manera:

Cluster	Localidades	%
Azul	75	59 %
Rojo	28	22 %
Verde	24	19 %
Total	127	100 %

Tabla 4.14: Clústeres CPV Yucatán.

A continuación se realiza la interpretación de clústeres con base en la dispersión de las variables originales, sus cuatiles poblacionales y el *embedding* espectral.

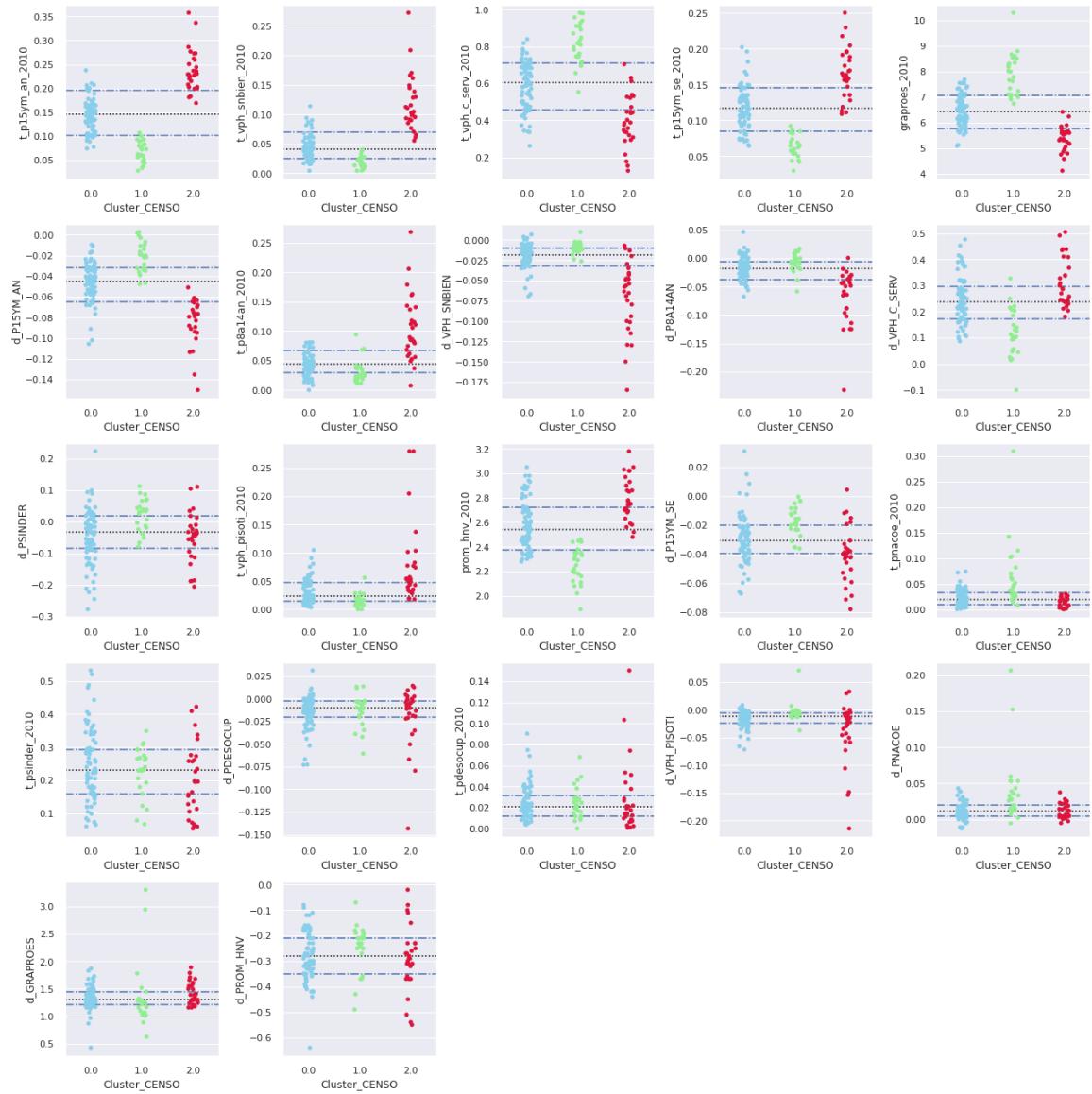


Figura 4.83: Dispersion clústeres CPV y cuantiles .25, .50 y .75 poblacionales.

En la Figura 4.84 se muestra la correlación lineal entre las variables y las dimensiones del *embedding*.

La dimensión 1 presenta una débil correlación con todas las variables y esto se debe a que corresponde al eigenvector del primer eigenvalor, cuyo valor siempre es 0. Sin embargo, valores ligeramente menos negativos se relacionan a localidades con mayor analfabetismo, viviendas sin bienes y viviendas con piso de tierra.

La dimensión 2 esta fuertemente asociada positivamente a localidades que en 2010 presentaban alto grado de escolaridad y alta cobertura de viviendas con servicios bási-

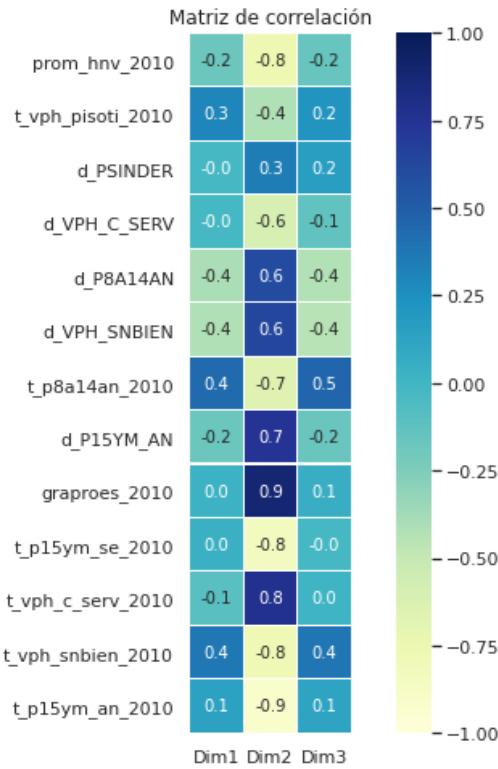


Figura 4.84: Matriz de correlación variables originales vs. *embedding*.

cos; por otro lado, valores negativos se asocian a las localidades que en 2010 contaban con altos niveles de analfabetismo, proporciones altas de personas sin escolaridad, viviendas sin bienes o con piso de tierra y un promedio alto de hijos.

La dimensión 3 es muy similar a la 1, se relacionan positivamente a localidades con mayor analfabetismo, viviendas sin bienes. Esta última dimensión aporta muy poca información, sin embargo, se decide mantenerla ya que al tener 3 clústeres es posible diferenciar los grupos de los extremos.

Clúster Azul:

- Este grupo contiene a la mayoría de localidades en Yucatán y representa el grupo con niveles intermedios en cuanto a bienestar social, ya que como se observó en el grafo este clúster se encuentra en la parte central y en un extremo se encuentra el clúster verde el cual tiene un mejor nivel de bienestar y en contra parte se encuentra el clúster rojo que presenta los peores indicadores.
- En las localidades de este clúster, en promedio el 60 % de las viviendas contaban

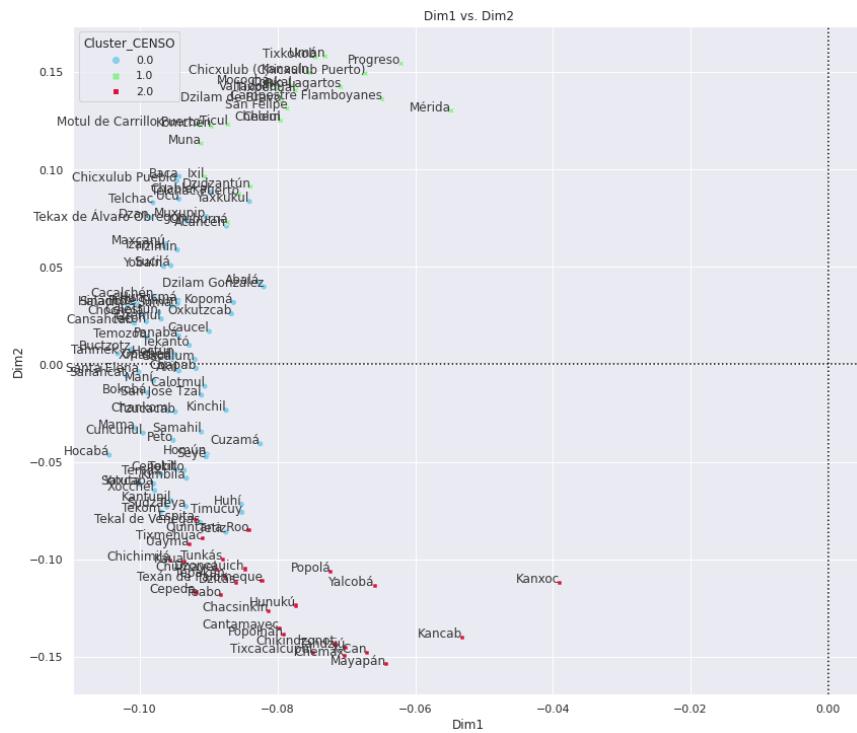


Figura 4.85: Dim1 vs. Dim2

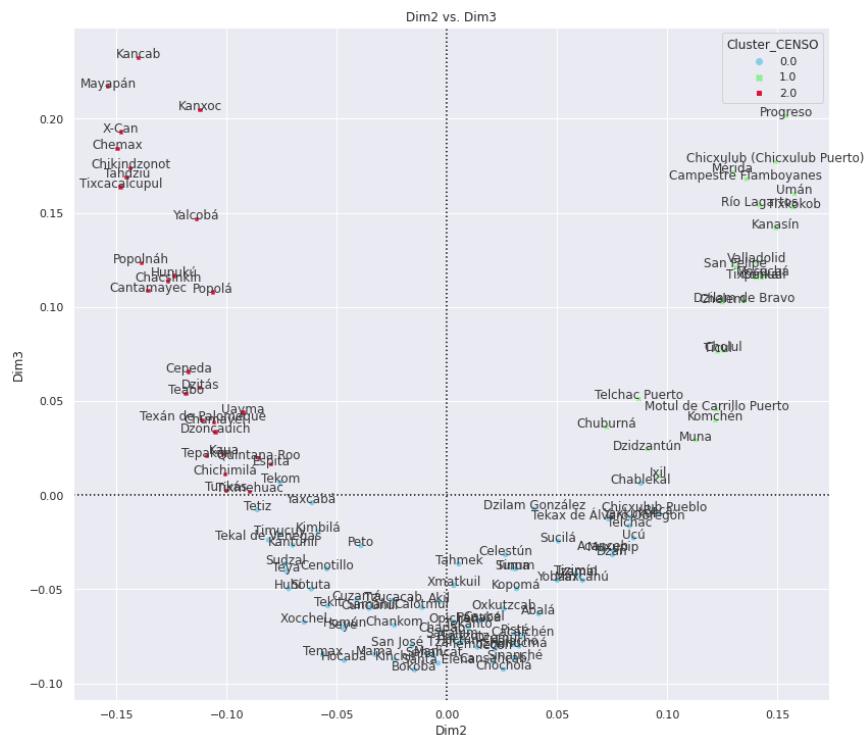


Figura 4.86: Dim2 vs. Dim3

con todos los servicios básicos en 2010 y en los últimos años ha mejorado ya que en 2020 en promedio el 84 % de las viviendas cuentan con servicios básicos.

- El grado de escolaridad, analfabetismo y personas sin escolaridad se encuentra dentro de los niveles medios respecto a el resto de localidades.

Las localidades en este clúster son las más neutrales a nivel entidad, han mejorado sus indicadores con el paso del tiempo pero hasta 2020 siguen manteniendo la misma posición intermedia en cuanto a bienestar social.

Cluster Verde:

- El grupo de localidades en este clúster presentan las menores tasas de analfabetismo y personas sin escolaridad desde 2010 y ha mantenido su posición hasta 2020. Presenta el grado de escolaridad promedio más alto del estado ubicándose en 9.2 en 2020.
- En las localidades de este clúster, en promedio el 81 % de las viviendas contaban con todos los servicios básicos en 2010, sigue mejorando ya que en 2020 en promedio el 93 % de las viviendas cuentan con servicios básicos.
- En los últimos años algunas de sus localidades han incrementado el porcentaje de habitantes nacidas en otra entidad, destacan Conkal, Cholul, Mérida, Chelem, Valladolid, entre otras.
- El promedio de hijos vivos en este clúster es el más bajo en la entidad.

Cluster Rojo:

- Las localidades en este clúster presentan las mayores tasas de analfabetismo y personas sin escolaridad desde 2010 y ha mantenido su posición hasta 2020. Su grado de escolaridad promedio es el más bajo del estado, se ubica en 6.8 años en 2020.
- En las localidades de este clúster, en promedio sólo el 40 % de las viviendas contaban con todos los servicios básicos en 2010, esto ha mejorado en los últimos

años ya que en 2020 en promedio el 71 % de las viviendas cuentan con servicios básicos.

- El promedio de hijos vivos en este clúster es el más alto en la entidad.

Con base en lo anterior, se propone el siguiente *ranking* por nivel de bienestar social, donde 4 estrellas es el máximo nivel y una estrella en nivel mínimo:

Cluster DENUE	Nivel de Bienestar	Categoría
Verde	★★★	A
Azul	★★	B
Rojo	*	D

Tabla 4.15: Ranking Clústeres DENUE Yucatán.

Una vez obtenido el *ranking* tanto del conjunto de variables del DENUE como del CPV es posible identificar qué localidades presentan un escenario favorecedor para los micro negocios y además cuentan con los mejores niveles de bienestar en la entidad. Dichas localidades se encuentran rankeadas de acuerdo a la suma total de estrellas, por lo que el mejor clúster será la categoría A-A de 8 estrellas en total y posteriormente la categoría A-B o B-A con 7 estrellas, y así sucesivamente.

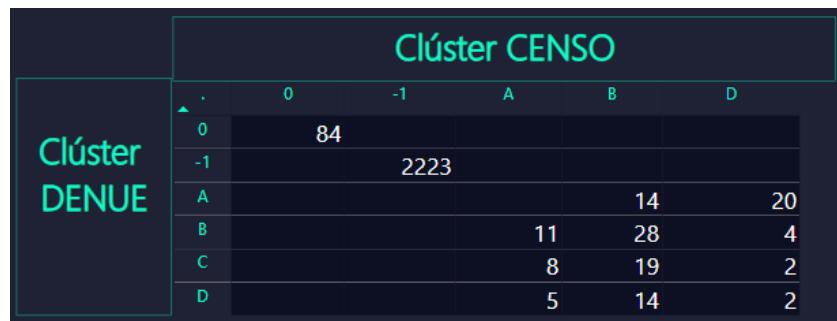


Figura 4.87: Localidades Yucatán por categoría.

Para el estado de Yucatán no se encontraron localidades de categoría A-A, por lo tanto, el siguiente grupo mejor posicionado es la categoría A-B que corresponde a 14 localidades que se identifican como zonas con potencial en el sector micro pero con niveles intermedios de bienestar social. Estas localidades se encuentran muy dispersas, contrario a lo que se observó en el estado de Nayarit y Nuevo León. Además, se observa

que contrario a Nayarit, las localidades de mejor potencial están más lejanas de la costa.



Figura 4.88: Localidades con mayor potencial de crecimiento micro y nivel intermedio de bienestar social Yucatán.

En estas localidades predomina, como en los otros estados, las tiendas de abarrotes y servicios de preparación de alimentos y bebidas, pero también hay fuerte presencia de la industria manufacturera textil. Además, se encuentran algunas variantes como el caso de Halachó, que es más cercano a costa, su enfoque es más hacia el comercio al por menor de ropa y accesorios, en Temozón el enfoque es más a la fabricación de muebles o la localidad Pisté que tiene presencia del sector manufacturero para fabricación de productos de madera, productos de cemento y concreto.

Por otra parte, el grupo de categoría B-A se conforma por 11 localidades que presentan un mejor nivel de bienestar social pero son localidades que antes de 2014 tuvieron su mayor crecimiento en el sector micro pero no continuaron creciendo y por lo tanto hoy se ubican en niveles promedio de crecimiento.

En general, las localidades de esta categoría tienen presencia en la industria alimentaria para la elaboración de productos como pan y tortilla y en el comercio de

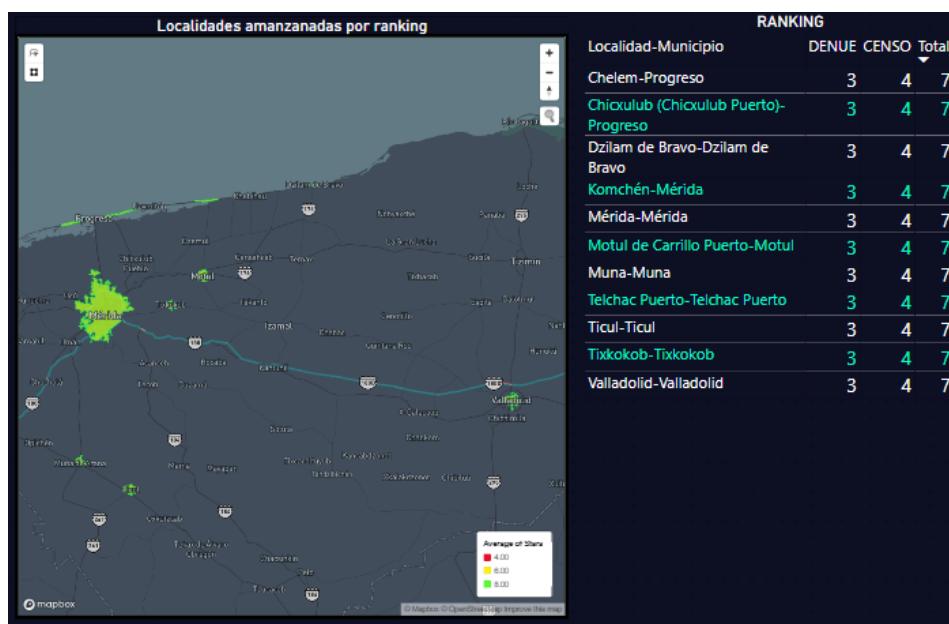


Figura 4.89: Localidades con mayor nivel de bienestar social y nivel medio de crecimiento del sector micro.

productos textiles, de calzado y de vestir. Sin embargo, algunas localidades como Chelém, ubicado en costa, proveen servicios de alojamiento temporal en forma de cabañas o villas y servicios de preparación de alimentos y bebidas

También hay algunas variantes de sectores como por ejemplo, en Dzilam el comercio de artículos de uso personal, artesanías y accesorios de vestir; en Mérida hay presencia de comercio al por mayor de materias primas para la construcción y en sí mismo el sector de construcción, en Muna encontramos manufactura textil, en Ticul industria manufacturera de fabricación de calzado y en Valladolid servicios de alojamiento temporal en forma de hoteles.

Por último, tomando como base las 127 localidades que pertenecen a alguna de las categorías (A,B,C,D) se obtuvo el siguiente mapa a nivel municipio iluminado de acuerdo al promedio de puntaje de las localidades que lo constituyen. Se puede observar que las zonas están muy heterogéneas a lo largo del estado, ya que encontramos zonas con mayor puntaje en costa centro y sur.

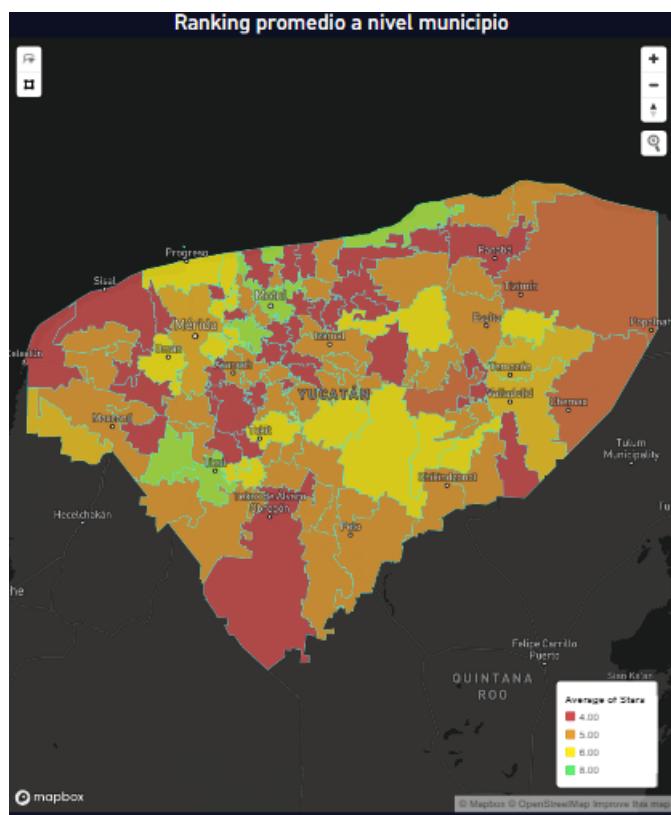


Figura 4.90: Ranking por municipio Yucatán.

Capítulo 5

Conclusiones y trabajo futuro

Las micro empresas forman parte fundamental de la economía mexicana, más de un tercio de la población se emplea por este medio y gran parte de los emprendimientos tienen origen en este sector.

Es importante destacar que, durante la pandemia por Covid-19, el sector micro registró porcentajes de muertes al mismo nivel que las PyME's, sin embargo, éste sobresalió por su mayor capacidad de regeneración.

En México, la recopilación de la información oficial de este importante sector la realiza el INEGI a través del DENU, no obstante, se observa que dicha información se actualiza después de periodos de tiempo muy largos; a pesar de que el DENU publica dos versiones por año, se observa que, en los periodos intermedios entre censos económicos, sólo un pequeño número de Unidades Económicas (UE) se dan de alta o actualizan su información. Así mismo, la información detallada acerca del nivel de bienestar social se obtiene por medio del Censo de Población y Vivienda, el cual se actualiza cada 10 años.

Dada la información pública disponible y el interés de identificar las zonas de mayor potencial por estado, se consideró conveniente aplicar un modelo de *clustering* con la finalidad de sintetizar la información de diversas variables y crear un *ranking* que reflejara las distintas condiciones que presentan las localidades.

En cuanto a la metodología utilizada:

- El *Clustering* Espectral significó una buena alternativa de solución, por medio de los grafos de similitud y el *embedding* espectral se obtuvo una representación visual de los objetos que residían en un espacio multidimensional, por tanto, el proceso de identificación de clústeres fue más sencillo e intuitivo.
- La técnica de selección de variables SPEC fue de gran utilidad para eliminar variables que no contribuían a identificar grupos en las observaciones, ya que éstas provocaban que las instancias produjeran grafos más compactos y se perdiera visibilidad de lo que realmente diferenciaba a los clústeres.
- La combinación del algoritmo de *Clustering* Espectral con la función de similitud RBF kernel y el parámetro local σ_i proporcionó buenos resultados, ya que permitió encontrar clústeres con formas desconocidas; además, el parámetro de escala local σ_i contribuyó a que el modelo fuera más flexible al adecuarse según la dispersión que presentaba cada clúster.
- La técnica SPEC y el algoritmo de *Clustering* Espectral presentaron resultados mucho más claros al aplicarlos por separado al conjunto de variables del DENUE y del Censo, en lugar de considerar un solo conjunto de variables.
- Los criterios Eigengap y Silhouette brindaron mayor claridad para medir la mejora de los clústeres cuando se eliminaron del modelo las variables de menor relevancia.

En cuanto a los resultados de aplicar la metodología desarrollada en los estados de Nayarit, Nuevo León y Yucatán, se obtuvieron las siguientes conclusiones:

- En general, las variables de mayor relevancia para la formación de los clústeres fueron las relacionadas a la antigüedad de las UE, la tasa de sobre-vivencia de las unidades dadas de alta en 2010, la proporción de UE que presentaron crecimiento en tamaño de personal, el grado de escolaridad, la tasa de analfabetismo, la proporción de personas sin escolaridad, el porcentaje de viviendas con servicios básicos y la proporción de habitantes nacidos en otra entidad.

- De los tres estados, el estado de Nayarit presentó grupos más diferenciados entre sí. En el caso de Nuevo León y Yucatán las observaciones eran más homogéneas de manera que fue más complicado encontrar los grupos, para estos dos estados fue favorable eliminar las variables de menor relevancia para poder visualizar de forma más clara los posibles grupos.
- Algunas de las relaciones observadas en las localidades fue que entre mayor era el grado de escolaridad en una localidad, menor era el promedio de hijos y mayor el porcentaje de habitantes nacidos en otra entidad. Las localidades de mayor potencial del sector micro registraron crecimientos de micro a pequeñas empresas y se asocian a localidades con todo tipo de antigüedades. Una cualidad a destacar es que localidades que se componen completamente de UE con antigüedades largas no implican necesariamente potencial de crecimiento, ya que se observó un mayor potencial en aquellas que se encuentran compuestas de forma equilibrada por UE de corta, mediana y larga antigüedad. Así mismo, localidades con mayor presencia de unidades de recién creación cuentan con los puntajes más bajos en el *ranking*.
- Los tres estados coinciden en cuanto a que *el Comercio al por menor* (tiendas de abarrotes) y el *Servicio de alimentos y bebidas* son los sectores de mayor presencia. Sin embargo, quitando estos sectores predominantes se puede observar que las zonas potenciales de cada estado presentan perfiles muy distintos; en Nayarit predomina el sector de servicios de alojamiento temporal (cabañas, villas y similares) y el comercio de productos textiles, de bisutería, accesorios de vestir y de calzado; en Nuevo León se tiene presencia de la industria manufacturera alimentaria, fabricación de productos de herrería, comercio al por mayor de materias primas agropecuarias, servicios de reparación mecánica y servicios personales como salones de belleza. En Yucatán hay presencia del sector manufacturero textil, comercio de ropa, fabricación de productos de madera, comercio de artesanía y fabricación de calzado.
- Las ubicaciones geográficas de las localidades con mejores puntajes de los tres

estados están distribuidos de manera distinta. En Nayarit, se ubican cercanos a la costa y a la frontera con Jalisco, en Nuevo León se ubicaron en zonas aledañas a la capital y con tendencia a expandirse hacia el norte y en Yucatán se encuentran muy dispersas alejadas de la capital y de la costa.

- Con un nivel de detalle a nivel localidad, en este trabajo se logró abarcar poblaciones tanto rurales como urbanas, ya que era de interés identificar zonas potenciales que no necesariamente se encontraran dentro de las áreas de mayor centralización. Sin embargo, si se desea llegar a un nivel de detalle más preciso, ya sea a nivel AGEB o manzana, la información pública del INEGI esta disponible pero solo para localidades urbanas, por lo tanto se perdería de vista al resto de localidades rurales. Un mayor nivel de detalle resulta adecuado para localidades de larga extensión territorial como es el caso de Monterrey o Apodaca, para estos casos es mucho más conveniente realizar el análisis a nivel AGEB o manzana, ya que dentro de esta localidad existen diferencias muy notables a muy cortas distancias.

Finalmente, como trabajo futuro, es de interés llevar a cabo un estudio enfocado en identificar qué sectores se relacionan a los existentes en las zonas de potencial crecimiento y que tienen nula presencia en la zona, como plantea Hausmann (2007) en el Atlas de Complejidad Económica, el crecimiento económico esta dado por la diversidad de productos en los mercados, lo cual se traduce en un incremento en el saber hacer colectivo. En este sentido, se buscaría fomentar dicha diversidad de productos.

Asimismo, es de interés buscar otras fuentes de información en el ámbito de información no estructurada, en la cual se tenga mayor continuidad en la actualización de los datos y por lo tanto un mayor número de observaciones a través de tiempo, lo anterior, con la finalidad de desarrollar un modelo de carácter predictivo, es decir, conocer la evolución de espacios geográficos para determinar hacia donde se dirige ese crecimiento.

Referencias

- Afzalan, M., y Jazizadeh, F. (2019). An automated spectral clustering for multi-scale data. *arXiv:1902.01990*.
- Aggarwal, C. C., y Reddy, C. K. (2014). *Data clustering algorithms and applications*. Taylor & Francis Group, LLC.
- Bernstein, M. (s.f.). The radial basis function kernel. *University of Wisconsin-Madison. Computer Sciences User Pages*. Descargado de <http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/svms/RBFKernel.pdf>
- Fleshman, W. (2019). Spectral clustering foundation and application. *Towards Data Science*. Descargado de <https://towardsdatascience.com/spectral-clustering-aba2640c0d5b>
- Halkidi, M., Batistakis, Y., y Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17.
- Hausmann, H. e. a. (2007). The atlas of economic complexity. maping paths to prosperity. , 51. Descargado de https://d1wqtxts1xzle7.cloudfront.net/30678659/HarvardMIT_AtlasOfEconomicComplexity_Part_I.pdf
- INEGI. (2018). El registro estadístico de negocios de méxico (renem) y el directorio estadístico nacional de unidades económicas (denue) . , 24. Descargado de <https://www.cepal.org/sites/default/files/presentations/presentacion-registro-estadistico-negocios-mexico-renem-directorio-estadistico-nacional-unidades-economicas-denue-inegi.pdf>
- INEGI. (2019a). Censos económicos 2019. Descargado de https://www.inegi.org.mx/programas/ce/2019/#Datos_abiertos

- INEGI. (2019b). Censos económicos 2019. resultados definitivos. , 59. Descargado de https://www.inegi.org.mx/contenidos/programas/ce/2019/doc/pprd_ce19.pdf
- INEGI. (2019c). Censos económicos 2019. resultados oportunos. , 21. Descargado de https://www.inegi.org.mx/contenidos/programas/ce/2019/doc/pro_ce2019.pdf
- INEGI. (2020a). Censos de población y vivienda 2020. principales resultados por ageb y manzana urbana. , *Segunda Edición.*, 41. Descargado de https://www.inegi.org.mx/contenidos/programas/ccpv/2020/doc/fd_agebmza_urbana_cpv2020.pdf
- INEGI. (2020b). Censos de población y vivienda 2020. principales resultados por localidad (iter). , *Segunda Edición.*, 43. Descargado de https://www.inegi.org.mx/contenidos/programas/ccpv/2020/doc/fd_iter_cpv2020.pdf
- INEGI. (2020a). Directorio estadístico nacional de unidades económicas denue interactivo 11/2020. documento metodológico. , 24. Descargado de https://www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825197414.pdf
- INEGI. (2020b). Estudio sobre la demografía de los negocios 2020. primer conjunto de resultados. , 21. Descargado de <https://www.inegi.org.mx/contenidos/programas/edn/2020/doc/EDN2020Pres.pdf>
- INEGI. (2021). Marco geoestadístico nacional. Descargado de https://www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/geografia/marcogeo/889463835615.pdf
- Liu, J., y Han, J. (2014). *Data clustering algorithms and applications*. Taylor & Francis Group,LLC.
- Luxburg, U. V. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4), 395–416.
- Ng, A., Jordan, M., y Weiss, Y. (2002). On spectral clustering: Analysis and algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems*. MIT Press., 14, 849-856.

- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20, 53-65.
- Shi, J., y Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine learning*, 22. No. 8, 888–905.
- Sreenivasa, S. (2020). Radial basis function (rbf) kernel: The go-to kernel. Descargado de <https://towardsdatascience.com/radial-basis-function-rbf-kernel-the-go-to-kernel-acf0d22c798a>
- Zelnik-Manor, L., y Perona, P. (2004). Self-tuning spectral clustering. *Advances in Neural Information Processing Systems*, 17:1601–1608.
- Zhao, Z., y Liu., H. (2007). Spectral feature selection for supervised and unsupervised learning. In *ICML '07: Proceedings of the 24th International Conference on Machine Learning*, New York., 1151-1157.

Apéndice A

Implementaciones y mapa interactivo

Parte de la contribución de este trabajo es la implementación en Python de la técnica de Selección de Variables Espectral (SPEC) y el algoritmo de *Clustering Espectral* propuesto por Ng y cols. (2002) que incorpora la variante del parámetro local σ_i en la función de similitud RBF kernel. El código generado puede ser consultado en el enlace de Google Colab:

[https://colab.research.google.com/drive/1z8TjqocS-kayVvVcAduFQpcsV1baK2VC
?usp=sharing](https://colab.research.google.com/drive/1z8TjqocS-kayVvVcAduFQpcsV1baK2VC?usp=sharing)

Así mismo, los mapas interactivos generados a partir de los resultados de este trabajo, pueden ser consultados en el siguiente enlace:

<https://app.powerbi.com/view?r=eyJrIjoiZDBiOWFkZWUtMWUyZSOONmViLTk3ZTgtYmNjOWQzYTUzMw==>



Figura A.1: Ejemplo Dashboard Nayarit

