VNIVERSIDAD NACIONAL
AVTNOMA DE
MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

MINERÍA DE DATOS VISUAL MEDIANTE LA EXPLORACIÓN DE REDES EMERGENTES

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIA
(COMPUTACIÓN)

PRESENTA:
RAÚL SIERRA ALCOCER

TUTOR PRINCIPAL:
DR. CHRISTOPHER STEPHENS STEVENS,
INSTITUTO DE CIENCIAS NUCLEARES

MIEMBROS DEL COMITÉ TUTOR:
DR. FERNANDO ARÁMBULA COSÍO,
CENTRO DE CIENCIAS APLICADAS Y DESARRROLLO TECNOLÓGICO

DR. PEDRO MIRAMONTES VIDAL,
FACULTAD DE CIENCIAS

MÉXICO, D. F. MARZO 2013

Resumen

La gran cantidad de datos a la que tenemos acceso en la actualidad nos ofrece nuevas oportunidades para entender nuestro mundo. La magnitud y complejidad de los datos necesitamos desarrollar nuevas técnicas para analizarlos. Uno de los principales obstáculos para entender un fenómeno es el número de factores que interactúan para su generación, ya que el número de interacciones posibles crece exponencialmente. Sin embargo, son estas interacciones las que en general nos interesa descubrir y comprender.

Cuando la información que tenemos acerca de algún fenómeno está dada por un conjunto grande de datos multivariados, el análisis exploratorio de estos datos es fundamental para extraer de estos la mayor cantidad de información posible. Una de las características del análisis exploratorio de datos es que el analista observe los datos desde distintas perspectivas. En este trabajo proponemos una metodología para analizar fenómenos desde la perspectiva de redes.

Las redes pueden ser utilizadas para representar visualmente sistemas de relaciones de manera intuitiva. Muestra de esto es que son utilizadas en distintos campos para tal fin. Aún así, las redes no son explotadas suficientemente, en particular en el análisis de datos espaciales. Por lo general, las redes se utilizan cuando los datos presentan explícitamente una estructura de red, como en una red social. En esta tesis presentamos una metodología que extiende el uso de redes a conjunto de datos donde no existe una relación explícita entre las variables. Nuestra metodología combina técnicas de análisis y minería de datos que integramos en un flujo de análisis cuya finalidad es

ayudar al analista a construir hipótesis a partir de la visualización de relaciones entre las variables. Para tal fin utilizamos estadísticas de diagnóstico, redes, visualización interactiva y mapeos auto-organizados. Todo esto lo complementamos con una discusión sobre los procesos que implica el análisis de datos en general.

La metodología de análisis que presentamos tiene cuatro etapas: (1) definir un modelo de coincidencia entre variables; (2) elegir el tipo de relaciones que nos interesan; (3) calcular estadísticas de correlación entre las variables a partir de las coincidencias y el tipo de relaciones; (4) construir y visualizar una red a partir de la estadística obtenida en el paso (3). En el trabajo ilustramos el potencial de la metodología con un ejemplo tomado del campo de la ecología. Finalmente, para explorar los patrones en la red desarrollamos un método para encontrar patrones de conectividad en una red utilizando mapeos auto-organizados (SOM).

Agradecimientos

Durante el desarrollo de esta tesis recibí el apoyo de muchas personas. Agradezco a todos por ayudarme a llegar a este punto.

Al Dr. Christopher Stephens, mi director de tesis, por su confianza, su entusiasmo y su compromiso con el proyecto. Ha sido mentor, colega y amigo. Espero que continuemos colaborando.

Al comité tutor, el Dr. Fernando Arámbula y el Dr. Pedro Miramontes. Por el tiempo que dedicaron, por sus aportaciones, porque su participación fue fundamental para que este trabajo de investigación se integrara.

Al Dr. David Rosenblueth y al Dr. Antonio Neme por haber aceptado ser sinodales. Gracias a sus observaciones esta tesis mejoró mucho, agradezco el tiempo que dedicaron a leerla.

A Lulú y a Diana por su apoyo en los procesos administrativos del posgrado. Porque cuando tuve algún problema inmediatamente me ayudaron a resolverlo con la mejor disposición.

También agradezco a las instituciones que ayudaron a que este trabajo fuera posible: el CONACYT, el IIMAS, el C3 y la UNAM.

A mi madre y a José Luis, por el cariño y el apoyo que me han dado.

A mi padre, que me transmitió su gusto por las matemáticas.

Finalmente, a Ale, mi esposa, por ser mi fuente de motivación. A ella y a nuestro hijo que está por llegar, les dedico esta tesis.

Tabla de Contenidos

1. Introducción **1**

2. Análisis de datos **11**

2.1. Análisis exploratorio de datos 14

2.1.1. Análisis visual de datos 15

2.1.2. Antecedentes 15

2.1.3. Desarrollos recientes 22

2.2. Inferencia 24

2.2.1. Experimentos controlados y estudios de observaciones 25

2.2.2. Prueba de hipótesis 26

2.2.3. Épsilon 30

2.3. Reducción de dimensiones 34

2.3.1. Sesgo vs Varianza 34

2.3.2. Análisis de componentes principales (PCA) 36

2.3.3. Escalamiento multidimensional (MDS) 39

2.3.4. Mapeos auto-organizados (SOM) 41

2.3.5. Métodos para datos no vectoriales 44

2.3.6. SOM para datos no vectoriales 44

3. Análisis de datos espaciales **47**

3.1. Diferencias con la minería de datos tradicional 49

3.2. Análisis visual de datos espaciales 50

3.3. Análisis de correlación espacial entre variables espaciales booleanas 51

3.3.1. Modelos de co-ubicación local 52

3.3.2. Épsilon para datos espaciales 54

3.3.3. Agregación de datos y el problema de la unidad de área modificable 55

3.4. Selección de rejilla para datos espaciales . . . 58
3.4.1. Exploración de los efectos de la resolución de rejilla . . . 58
3.4.2. Optimización de la resolución de la rejilla . . . 66
3.4.3. Optimización del número de coincidencias . . . 67
3.4.4. Resultados . . . 68

4. Redes para el análisis de datos **73**
4.1. Aplicaciones de redes . . . 75
4.1.1. Redes espaciales . . . 75
4.1.2. Modelos gráficos . . . 77
4.1.3. Redes en procesamiento del lenguaje natural . . . 77
4.1.4. Patrones de co-ubicación . . . 79
4.2. Redes y análisis de datos espaciales . . . 81
4.3. Redes inferidas para datos espaciales . . . 84
4.3.1. Caso de estudio . . . 88
4.3.2. Aplicación . . . 90
4.4. Agrupamiento de nodos en redes . . . 96
4.5. Medidas de similitud entre nodos . . . 99
4.6. SOM para agrupamiento de nodos en redes . . . 102
4.6.1. SOM para agrupamiento de nodos en redes dirigidas y pesadas . . . 103
4.6.2. Algoritmo . . . 107
4.6.3. Evaluación . . . 109
4.6.4. Resultados del SOM para agrupamiento de nodos . . . 111

5. Conclusiones y perspectivas **117**
5.1. Trabajo futuro . . . 121

Bibliografía **123**

Índice de figuras

2.1. Mapa del centro de Londres que muestra los cúmulos de casos de cólera durante la epidemia ocurrida en 1854. Cada caso se muestra como una línea negra en el domicilio donde ocurrió. Se puede observar cómo los casos se apilan alrededor de un hidrante en Broad Street y Little Windmill Street. John Snow, 1854. Fuente: http://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg 17

2.2. Diagrama de área polar que muestra estadísticas de las bajas en el Ejército del Este durante la guerra de Crimea entre abril de 1884 y abril de 1856. Cada color representa en área medida desde el centro, el número de muertos en un mes. La secciones azules representan muertes prevenibles, las secciones en rosa representan muertes por heridas de guerra, y las secciones en negro muertes por otras causas. Florence Nightingale, 1858. Fuente: http://es.wikipedia.org/wiki/Archivo:Nightingale-mortality.jpg 18

2.3. Carta figurativa de las pérdidas en hombres de la Armada francesa durante la campaña de Rusia 1812-1813. La barra café representa las tropas del ejército francés en dirección a Moscú, mientras que la negra representa las tropas en retirada. En ambas barras un milímetro de ancho (en el original de 55.88 x 50.8 cm) representaba 6000 hombres. Charles Minard, 1869. Fuente: http://en.wikipedia.org/wiki/File:Minard.png 19

2.4. Visualización de treemap de un sistema de archivos usando la aplicación GrandPerspective. El espacio completo representa un directorio y los rectángulos que lo subdividen los subdirectorios o archivos contenidos en el directorio. El color de un rectángulo indica el tipo de archivo, y su área el porcentaje del total de memoria que ocupa el archivo. 23

2.5. $\epsilon_{x,y}$ es la distancia entre μ_{S_X} y $\mu_{S_{X|Y}}$ en desviaciones estándar de S_X 33

2.6. Ejemplo de modelos que aproximan un mismo conjunto de datos, y donde la elección clara es el modelo lineal. 35

3.1. Co-ocurrencias y ε entre dos variables espaciales booleanas como funciones de la resolución de rejilla. La gráfica azul corresponde a variables independientes y la roja al caso donde X depende de Y. 69
3.2. Número esperado de coincidencias entre dos variables espaciales independientes X, Y como función del número de celdas. 70
3.3. Comparación entre el número esperado de coincidencias entre X, Y y el número esperado de celdas ocupadas por X. 70
3.4. Cada columna corresponde a una especie de mamífero y cada renglón a una resolución. 71
3.5. Búsqueda del máximo de coincidencias para 8 especies. 71
3.6. 72

4.1. Ejemplo de red bayesiana de cuatro variables, esta red implica por ejemplo $P(A, B, C, D) = P(A)P(B \mid A)P(C \mid A)P(D \mid B, C)$ 78
4.2. Tipos de red que pueden definir patres de co-ubicación. 80
4.3. Distribuciones geográficas de *Lutzomyia* cruciata y Peromyscus leucopus. 89
4.4. Red vector–reservorio. 91
4.5. Resultado de eliminar aristas que no cumplen $\varepsilon \geq 6$. 92
4.6. La topología de la red está ligada a la distribución geográfica de las características espaciales booleanas. 94
4.7. Una arista gruesa entre dos nodos pequeños muestra una relación fuerte entre dos especies poco comunes. 95
4.8. Ejemplo de una red con 4 grupos de nodos que se conectan igual 104
4.9. Grupos de nodos en 4.8 104
4.10. Interfaz del sistema de sustratos semánticos [4] 105
4.11. 114
4.12. 115
4.13. Red épsilon de mamíferos con especies del género *Lutzomyia* y resultado de análisis de conectividad usando el SOM para nodos 116

Índice de tablas

3.1. X, Y independientes. Tabla que muestra los resultados de corridas con resoluciones cada vez más finas, con 500 puntos de X y 100 de Y, ambas con distribución aleatoria uniforme 61
3.2. X dependiente de Y. Tabla que muestra los resultados de corridas con resoluciones cada vez más finas, con 500 puntos de X y 100 de Y con distribución uniforme. Donde la distribución de X depende de la de Y. [EXPLICAR COMO] 62

Capítulo 1

Introducción

En las últimas dos décadas ha habido un crecimiento espectacular en la cantidad de datos disponibles, y nuestra capacidad para generar y distribuir datos crece rápidamente, por lo que se requieren grandes esfuerzos para que nuestra capacidad de usar estos datos crezca a la par. En 2002, un estudio del Departamento de Ciencia de la Información de la Universidad de Berkeley titulado "How much information?", obtuvo estimados sobre las tendencias que siguió la producción, el almacenamiento y el consumo de datos. En dicho estudio se observó una tendencia reveladora: entre 1999 y 2002 la cantidad de datos almacenados aumentó a una tasa de alrededor de 30 % anual. Aunado a esto, se estimó que la superficie de la Web (páginas estáticas) contenía en ese momento aproximadamente 127 terabytes, lo cual equivale aproximadamente a 127 millones de libros o a un librero de 399 km de largo. Pero esto es sólo la punta del iceberg, ya que no se considera el contenido de todas las bases de datos accesibles por internet a través de sitios dinámicos ni a través de API's. Con esta gran cantidad de datos sobre nuestro mundo, el reto es desarrollar herramientas que nos permitan convertir esos datos en conocimiento.

Al mismo tiempo que ha aumentado la capacidad de registro, almacenamiento y distribución de datos, también ha aumentado nuestra capacidad para analizarlos. Esto tiene implicaciones en la

manera en que podemos hacer ciencia, e implica nuevos paradigmas que se han venido desarrollando en las últimas décadas. Mientras que en el pasado se comprobaban hipótesis a partir de mediciones específicas en experimentos controlados, actualmente ese método no siempre es viable. Ahora contamos con muestras gigantes de datos que involucran a su vez grandes cantidades de variables, sin embargo, todos estos datos han sido registrados con diversos propósitos –en ocasiones sin propósito explícito alguno–. Esta nueva condición donde los datos ya se han recolectado, pero no forzosamente para el fin con el que son estudiados, implica que se requieren buscar nuevos métodos que nos permitan sacar la información que contienen. Para esto requerimos explorarlos desde una multiplicidad de perspectivas que nos permitan entender su naturaleza. Esto presenta nuevos retos y nuevas oportunidades.

Un objetivo en el análisis de datos es buscar correlaciones entre las variables que nos expliquen algo sobre el comportamiento de un sistema. El poder de cómputo no sólo nos da la habilidad de procesar datos en masa para realizar cálculos, también nos da la flexibilidad de explorarlos con mayor libertad de lo que se podía en el pasado. Sin este poder de cómputo, buscar la correlación entre dos variables puede ser una tarea titánica, por lo que habría que ser mucho más selectivo y usar conocimiento a priori para restringir el espacio de búsqueda; el problema es que al hacer esto podríamos eliminar patrones útiles en los datos, pero que descartamos por no estar en concordancia con nuestras hipótesis sobre el fenómeno que produjo el conjunto de datos. Actualmente, podemos darnos el lujo de explorar relaciones que parecieran improbables sin tener una razón de peso más que una simple sospecha. Estas nuevas capacidades son prometedoras, pero no son la panacea, se requieren nuevas metodologías que ayuden a resaltar las relaciones escondidas en los datos, y nuevas representaciones de los datos que nos permitan procesarlos cognitivamente.

El análisis de grandes conjuntos de datos es un problema que presenta muchos retos. Dos problemas importantes son: el número de variables y el objetivo con que esos datos fueron recolectados. El primer problema se refiere al tamaño del espacio de exploración; es como si quisiéramos reconstruir

la orografía del mundo a partir de unos cuantos registros de altitud en la superficie terrestre. El segundo problema se refiere a que una gran parte de los datos a los que tenemos acceso no fueron registrados con nuestro problema específico en mente, lo cual resulta en una diferencia importante con respecto a métodos tradicionales de la ciencia, donde se toman mediciones controladas en laboratorios con un objetivo bien definido, o al menos las observaciones que se registran fueron diseñadas de acuerdo con las hipótesis del estudio.

Para atacar el primer problema generalmente simplificamos el espacio de exploración mediante técnicas de reducción de dimensiones (p. ej. con selección de características) y al mismo tiempo simplificamos las propiedades del sistema, aunque esta simplificación no esté apegada a la realidad (p. ej. asumir independencia de variables). Esta simplificación nos facilita la implementación de metodologías de exploración de datos. El análisis exploratorio de datos después nos permite refinar nuestras asunciones sobre las propiedades del sistema y construir nuevas hipótesis.

La exploración de datos no sólo es útil para mejorar nuestras hipótesis sobre el fenómeno que generó los datos, también es útil en el contexto del segundo problema. Cuando los datos fueron recolectados con objetivos ajenos al de nuestro estudio hay que entender cómo fueron recolectados; qué implicaciones tiene el método de recolección en las distribuciones de las variables; o qué errores de captura pueden haber en los datos. Es importante entender esto para limpiar lo más posible el conjunto de datos y para entender que debilidades pueden tener nuestras conclusiones.

Un término que surgió para identificar este tipo de problemas de análisis de datos es minería de datos. En la literatura podemos encontrar distintas definiciones de minería de datos. Pero en general, la minería de datos se define por su objetivo: el descubrimiento de patrones útiles y no evidentes en, usualmente grandes, conjuntos de datos. Lo cual es, en esencia, análisis de datos con la implicación de que es análisis de muchos datos, datos que sin computadoras no podríamos analizar en su totalidad. A lo largo de este trabajo intercambiaremos análisis de datos con minería de datos libremente.

La minería de datos es un proceso que, en términos generales, involucra los siguientes pasos: preparación de datos, exploración de datos, y creación de modelos para apoyar la toma de decisiones [29]. Pero estos pasos difícilmente se pueden realizar de forma automática. Es importante que las capacidades cognitivas del ser humano estén integradas en el proceso de minería de datos y en particular es fundamental en la exploración de dicho datos.

Si queremos lograr la inserción del analista en el proceso, es importante encontrar representaciones visuales de los datos que ayuden a la comprensión de su estructura. De esto se ha encargado la estadística por siglos, desarrollado un gran número de técnicas. A partir de finales del siglo pasado, gracias al desarrollo computacional, y con la aparición de grandes bases de datos multivariados, se ha buscado escalar estas técnicas para aplicarlas a conjuntos de datos masivos. Al mismo tiempo se han buscado soluciones desde áreas como el aprendizaje de máquinas que permiten desarrollar métodos mejor adaptados para este tipo de análisis. Además se ha reafirmado la importancia de encontrar representaciones visuales de los datos que ayuden a la comprensión de su estructura. La importancia de la representación visual es evidente si pensamos en términos de las capacidades de nuestros sentidos y su papel en nuestra interacción con el mundo. Nuestro sistema visual es nuestro sentido con mayor ancho de banda: es con el que podemos procesar la mayor cantidad de información, y está optimizado para la detección de patrones.

De nada sirve contar con grandes cantidades de datos si no tenemos las herramientas adecuadas para explorarlos. El diseño de visualizaciones de datos no es una cuestión estética. Se trata de codificaciones visuales basadas en un marco teórico proveniente de: ciencias cognitivas, investigación en interacción hombre-máquina, y de estudios de caso que han servido para entender las tareas que tienen que resolver los analistas –ya sean expertos de área, o analistas casuales–. La manera en que se usan los colores, las formas y las herramientas de interacción son todos parte de lograr una visualización exitosa. En el artículo [1], los autores proponen que todos somos analistas espacio-temporales. Creemos que esta idea se extiende a: todos somos analistas de datos.

Este trabajo presenta un flujo de análisis que esperamos resulte intuitivo y facilite el análisis de correlaciones en conjuntos de datos multivariados. El cual permita al analista la construcción de hipótesis y su refinamiento, y ayude a definir el camino para la construcción de modelos tanto predictivos como descriptivos. De manera abstracta un flujo de análisis se compone de tres partes: datos, tareas y herramientas (se pueden pensar como el qué, el para qué, y el cómo). En el tipo de flujos con el que se trata en este trabajo los datos son datos con un gran número de variables, la tarea es entender cómo se relacionan las variables entre sí, y las herramientas son inferencia estadística y visualización de redes.

El objetivo del flujo de análisis que planteamos es la comprensión de un fenómeno mediante la exploración del conjunto de datos utilizando una combinación de métodos de inferencia y visualización. En particular consideramos que es útil cuando tenemos una gran cantidad de datos, pero poco entendimiento acerca de lo que representan. Otra de las características de nuestra metodología es que está diseñada para ser intuitiva en comparación con muchas herramientas de análisis que son útiles sólo para especialistas o son cajas negras que sirven en labores de predicción y toma de decisiones pero no ayudan en la comprensión del fenómeno.

Antes de continuar, creímos necesario hacer una pausa para una reflexión corta sobre correlación y causalidad. Actualmente para gran parte de los datos a los que tenemos acceso no tenemos control sobre cómo y cuándo se recolectan. Esta situación complica más un ya desde antes complicado debate. Este debate se refiere a cuándo podemos hablar de causalidad en una relación entre variables. En el caso de datos dónde no tenemos control sobre la generación de estos se dice que no se puede demostrar causalidad debido a que no podemos manipular una variable para demostrar que ejerce una influencia directa en otra variable. Esta en realidad es una discusión que llega al ámbito de la filosofía y que no buscamos resolver en esta tesis. Esto, sin embargo no implica que las conclusiones no sean útiles, ya que aunque las relaciones sean al nivel de correlación, esta información sí nos permite construir modelos predictivos, nos da elementos para la recolección de nuevos datos que

refinen los modelos, y nos ayuda en la toma de decisiones. Sólo debemos mantener en mente que correlación no implica causalidad.

En esta investigación ha sido difícil encontrar una terminología adecuada debido a que los temas son comunes a varias disciplinas, y cada una tiende a desarrollar una terminología que se apegue a sus términos y prácticas particulares. En el libro 'Exploratory Analysis of Spatial and Temporal Data' [2], los autores proponen una terminología que se deslinda del origen de los datos y de campos particulares y se apega al proceso de análisis de datos de manera abstracta. Esta terminología, en particular, es útil para hablar de la estructura de un conjunto de datos, en especial en lo relacionado a los componentes de referencia y los componentes de características de un conjunto de datos, conceptos que explicaremos y utilizaremos más adelante.

Veremos el análisis de datos como un proceso que involucra datos, tareas y herramientas. Los datos están compuestos por unidades que conocemos como registros; los registros comparten una estructura determinada por componentes; y cada componente representa una variable o característica. Por ejemplo, en un censo de población cada registro corresponde a una persona encuestada, y el registro contiene: dirección, escolaridad, edad y sexo. A partir de este conjunto de datos, uno puede plantearse distintas preguntas, por ejemplo, ¿cuál es la distribución de persones menores de 18? o ¿cuál es el nivel de escolaridad más común en una zona?, en estas preguntas implícitamente se define el rol de los componentes, el rol que juegan puede ser de referencia o de característica. Los componentes característicos dependen de los componentes de referencia, por ejemplo, la dirección de una persona es un componente de referencia y las características de esa persona son componentes característicos, dada una dirección podemos preguntarnos cual es el perfil de la persona (escolaridad, edad, ingreso, etc). En general los componentes de referencia son de los siguientes tipos: espacial, temporal y de población (donde población se refiere a cualquier conjunto de objetos sobre los que se hayan hecho las observaciones)[2].

El objetivo del análisis de datos a grandes rasgos es llegar a que el analista infiera (es decir, sa-

que una consecuencia o deduzca algo a partir de los datos) las respuestas a preguntas sobre un fenómeno, una de las herramientas para conseguir este fin es la inferencia estadística.

La inferencia estadística es el proceso para obtener conclusiones acerca de datos provenientes de variables influenciadas por procesos aleatorios. De acuerdo con el punto de vista bayesiano, en realidad no estamos hablando de procesos aleatorios, sino de variables de las que tenemos algún grado de incertidumbre sobre su comportamiento, debido a que carecemos de toda la información, en términos de que estamos lidiando con un problema de incertidumbre, el planteamiento bayesiano resulta más natural. Pero sea cual sea la forma de verlo, nuestro objetivo es amplificar las capacidades de inferencia del usuario apoyándolo con una combinación de métodos de inferencia estadística y visualización.

El procedimiento de inferencia requiere que se establezca un modelo de referencia, el cual asume ciertas propiedades (hipótesis) sobre las variables generadoras de los datos. En general, esto implica algún tipo de simplificación del sistema, ya sea por ignorancia (falta de datos), porque se considera que algunos factores tienen poca relevancia, o porque resulta demasiado complejo considerar todos los detalles de información que se tienen sobre el fenómeno. Esto nos lleva a un problema importante en la inferencia estadística. Es común que se requiera hacer un sacrificio entre la integración de todos los datos y la construcción de modelos que extraigan las propiedades necesarias para reproducir el comportamiento del fenómeno. El equilibrio entre varianza y sesgo. Si uno intenta que el modelo replique exactamente los datos se elimina el sesgo y se incrementa la varianza del modelo, por el contrario, si se simplifica el modelo, por ejemplo, al asumir un modelo lineal en vez de un modelo no lineal, se reduce la varianza, pero se puede incrementar el sesgo debido a que se deja fuera información en los datos. Si la varianza de los datos se reproduce exactamente, es decir, el modelo se ajusta perfectamente a todos los datos originales, los modelos tienden a padecer de sobreajuste con lo cual se pierden capacidades de predicción o de integrar nuevos datos. Del mismo modo, un modelo que simplifica demasiado el fenómeno, tiende a sacrificar la integración de la va-

rianza que existe en los datos, por lo que tendrá poca capacidad para describir el fenómeno, debido a que tiende a distanciarse del valor de las datos y pierde información sobre patrones regulares que pueden existir en ellos. Todo esto implica que en inferencia estadística habrá que modular la complejidad de las hipótesis que asume el modelo para obtener modelos que describan suficientemente bien los datos que tenemos y tengan la flexibilidad para integrar nuevos datos o para predecir eventos futuros.

La característica principal de nuestro flujo de análisis es que estamos interesados en sistemas de interacciones, donde las interacciones se dan entre objetos, características o entes, los cuales han sido observados en un conjunto de eventos. Donde las interacciones son desconocidas ya sea parcial o totalmente y parte de la tarea consiste en inferirlas. La razón por la cual nos interesa plantear el problema en términos de interacciones es la premisa de que el universo está de algún modo construido a partir de interacciones. Estas interacciones se dan de forma local en tiempo y lugar, para construir interacciones a mayor escala que se perciben como interacciones a distancia. Estas interacciones inferidas nos permiten estudiar los sistemas como redes que podemos explorar visualmente.

Aunque las visualizaciones de datos tienen una larga historia, no fué sino hasta los años 80 que la visualización de información empezó a generar interés como un campo desde el punto de vista científico. Una muestra de la relevancia que ha cobrado son la serie de conferencias de formación reciente en comparación con otros campos que existen actualmente: (Visweek en EEUU –que comprende: IEEE Visualization (Vis), IEEE Information Visualization (InfoVis), y IEEE Visual Analytics Science and Technology (VAST), Biological Data Visualization (Biovis, 2011 fue el 1er Symposium), International Workshop on Visualization and Collaboration (VisualCol 2012); Eurovis en Europa; IEEE Pacific Visualization en Asia). El tipo de temas que aborda la visualización de información es por naturaleza multidiciplinario ejemplo de esto son las áreas a las que pertencen algunos de los exponentes más reconocidos en esta área, como: Stephen Few (Business Intelligence), Edward Tufte (Diseño), Colin Ware (Percepción), Ben Shneiderman (Interacción Hombre Máqui-

na). Finalmente, también se puede ver esto en el surgimiento de revistas dedicadas al tema o el espacio que revistas de más tradición le dedican a este tema. La revista Information Visualization, Palgrave, se creó en el 2002; por otro lado la revista Transactions in Computer Graphics and Visualization, IEEE, publica constantemente artículos sobre visualización de información; la revista International Journal of Geographical Information Science, Taylor and Francis, es otro ejemplo de revista donde ocupan un espacio importante artículos relacionados con métodos de visualización de datos espaciales. Una muestra de la actividad que hay en el campo y de sus orígenes recientes es la creación de nuevos términos que siguen buscando como definir mejor el área: Visual Analytics y Geovisual Analytics, se empezaron a utilizar apenas en el 2005, [53].

Como ejemplos de visualización de datos en estadística clásica tenemos, por mencionar algunos, los histogramas y los diagramas de dispersión. Los primeros nos sirven para analizar la distribución de una variable, mientras que los segundos sirven para analizar la relación de dependencia entre dos variables. Pero, cuando se tienen más de dos variables y miles de registros que se quieren explorar para entender cómo se relacionan las variables entre si, el primer problema reside en cómo representar estos conjuntos de relaciones que no se pueden visualizar directamente en 2 o 3 dimensiones. Para lograr esto, se requiere algún tipo de proyección que respete lo más posible la estructura de las relaciones, y que mapee los datos a dos o tres dimensiones.

Las redes ofrecen una estructura que ha sido reconocida en muchos campos por ser un tipo de representación visual útil, debido a que permiten representar sistemas de relaciones complejos de forma intuitiva. Las redes, además, proveen de una teoría matemática sólida para su análisis algorítmico y estadístico, principalmente en teoría de gráficas y más recientemente por resultados obtenidos en el estudio de sistemas complejos. En particular, en ciencia de la información geográfica hay una larga tradición en análisis de redes para el estudio de redes espaciales [13]. Las redes espaciales son por lo general redes definidas por un conjunto de nodos que representan localidades y un conjunto de aristas que representan conexiones entre estas localidades, estas conexiones pueden ser conexiones

estructurales como carreteras o por flujos como rutas de aviones entre aeropuertos. En estos casos, las redes son utilizadas como una visualización extra para explorar las propiedades topológicas y geométricas de redes de infraestructura [60], o para entender la dinámica de flujos entre localidades [47]. Aunque las redes son usadas consistentemente para redes espaciales, su uso no se ha extendido a la minería de datos espaciales en general, cuando los datos no conforman explícitamente una red.

En los capítulos siguientes revisaremos los temas mencionados en esta introducción y presentaremos el desarrollo de una metodología que comprende construcción de un modelo de exploración a partir de la simplificación de las reglas (asumir independencia de variables), de selección de variables como método de reducción de dimensiones y visualización para el análisis exploratorio interactivo.

Capítulo 2

Análisis de datos

Para desarrollar una teoría sobre el análisis de datos hay que definir a qué nos referimos con "análisis de datos", para esto es necesario ubicar su objetivo. La finalidad de todo análisis de datos es contestar preguntas acerca de algún fenómeno especifico a partir de un conjunto de datos relacionado con este. Lo interesante es que aunque pareciera que hay un acuerdo general, si vemos las distintas definiciones que se encuentran en la literatura, nos encontramos con que esto no parece estar tan claro. Por ejemplo, una definición de análisis de datos podría ser: 'es el proceso de computar varios resúmenes y valores derivados de una colección de datos, es un proceso iterativo, cuyo objetivo es descubrir' [28]. En [2], los autores comparan esta definición con la que aparece usualmente en ciencia de la información geográfica: 'Un proceso para buscar patrones geográficos en los datos y relaciones entre características'. Seguramente si uno se dedica a buscar encontrará muchas más definiciones. Si queremos llegar a una definición satisfactoria para establecer un marco teórico sobre el análisis de datos, debemos fijarnos en lo que hay en común entre estas definiciones, ya sea explícita o implícitamente. Los autores de [2] hacen una reflexión en ese sentido y proponen que cualquier análisis de datos se conforma de los siguientes elementos:

1. Formular preguntas

2. Elegir métodos de análisis

3. Preparar los datos para aplicar los métodos

4. Aplicar los métodos

5. Interpretar y evaluar los resultados obtenidos

Hay que recalcar que el proceso es iterativo, es decir, el análisis implica regresar a la generación de nuevos modelos derivados una y otra vez, para incluir los descubrimientos que se hayan logrado a partir de pases previos. Podríamos decir que un objetivo del análisis de datos es obtener simplificaciones de estos. Simplificaciones en el sentido de que reducen las variables que interactúan en el fenómeno, acotándolas a las que parecen describir el sistema, partiendo de hipótesis sobre el comportamiento de estas variables, en ocasiones implícitamente. En cierto sentido la forma en que percibimos un fenómeno es siempre una simplificación. La temperatura que sentimos o cuando vemos un objeto moverse son en realidad sistemas complejos que registramos como un todo sin necesidad de saber como funciona el sistema de partículas y energía que los genera, y para efectos prácticos (más allá de que podamos o no podamos percibirlos con todo el detalle) de este modo es como nos conviene registrarlos porque a ese nivel podemos reaccionar ante ellos y tomar decisiones. La idea de simplificación del sistema es fundamental porque los seres humanos tenemos límites en cuanto a la cantidad de datos que podemos procesar mentalmente. Por esta razón requerimos simplificar los sistemas en su número de variables e interacciones, y sintetizar esos datos en información sobre tendencias o patrones.

La simplificación de la que hablamos no es arbitraria, ya que las formas en que es útil simplificarlo dependen de las características del sistema. Es decir la simplificación particular que se construye para un análisis es producto de la relación entre nuestras preguntas y las características del fenómeno. Eso mismo requiere a su vez de un análisis para detectar cuáles son las simplificaciones

adecuadas, recalcando que “adecuadas” se refiere tanto al tipo de simplificaciones que permite el fenómeno como a los objetivos de la investigación. En general, esta simplificación implica también una elección en el grosor de granularidad (coarse graining) que se aplica a los datos, ya que un estudio particular difícilmente analiza el fenómeno a todos los niveles de detalle posibles.

Por lo tanto, la lista anterior requiere algunas modificaciones. Una tarea que falta es que necesitamos buscar y decidir qué datos vamos a utilizar; también marcar que el proceso es iterativo; y que el medio para construir interpretaciones de los datos es la simplificación. Así que, dado un fenómeno y un conjunto de datos, el análisis sigue los siguientes pasos:

1. Formular (o refinar) preguntas
2. Elegir métodos de análisis
3. Preparar los datos para aplicar los métodos
4. Aplicar los métodos para obtener una simplificación del sistema
5. Interpretar y evaluar los resultados obtenidos(Volver a los pasos 1 y 2 tantas veces como sea necesario)
6. Obtener nuevos datos (Si es posible, y volver al paso 1)

Esta definición es a su vez una simplificación que captura los elementos que consideramos necesarios para nuestros fines. Un último punto que debemos recalcar es que la elección de datos es un primer filtro ineludible. Uno elige los datos de acuerdo con las preguntas que uno busca responder acerca del fenómeno. Esto es, ante el mismo fenómeno –por ejemplo, el descenso de una canica en una pendiente– podemos formular distintas preguntas, y de acuerdo con estas preguntas haremos una selección de los datos que consideramos necesarios. Incluso teniendo los mismos datos las preguntas que buscamos responder determinan las características del problema de análisis y el tipo de simplificaciones que requeriremos.

2.1. Análisis exploratorio de datos

John Tukey (el matemático estadounidense que desarrollo el algoritmo de la transformada rápida de Fourier) en un esfuerzo por enfatizar la importancia en el análisis de datos de la construcción de hipótesis, selección de herramientas estadísticas, y la recolección de datos, definió el Análisis Exploratorio de Datos [55]. El análisis exploratorio de datos es una filosofía mas que un conjunto de técnicas específico. Se trata de construir hipótesis a partir de los datos, en vez de buscar probar una hipótesis elaborada de antemano. Para lograr la construcción de hipótesis lo primero que debe hacer el analista es familiarizarse con los datos, y para ello en general se recurre a visualizaciones de los datos o de propiedades estadísticas de los datos. Encontramos en el libro "Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach" [2] fragmentos que elaboran sobre esta lógica (en traducción libre):

"La mayor parte del análisis exploratorio de datos es gráfico por naturaleza con algunos métodos cuantitativos. La razón por la que se depende de forma tan marcada es que el rol del AED es explorar teniendo la mente abierta, y las gráficas dan al analista poder sin paralelo para hacer esto". Más adelante acerca de los objetivos del análisis exploratorio de datos nos dicen que "la meta principal del AED puede ser vista en general como la construcción de un patrón apropiado a partir del comportamiento general definido por nuestro conjunto de datos completo".

El trabajo de Tukey presenta una lista de gráficas para la visualización de estadísticas de los datos de manera informal, ya que Tukey en su afán de presentar una forma de búsqueda en los datos deslindada de estructuras predefinidas no creía en la formalización de su trabajo, al menos no hecha por él. Mientras Tukey desestimaba la formalización de técnicas de análisis de datos, Jacques Bertin publicó "Semiótica de las gráficas"[6], la primera formalización teórica sobre el proceso de codificación de la información en elementos visuales. Dicho trabajo es considerado por muchos como el tratado fundacional de la visualización de información como disciplina científica. A con-

tinuación presentamos un breve recuento del desarrollo de la visualización como herramienta de análisis de datos.

2.1.1. Análisis visual de datos

2.1.2. Antecedentes

Los orígenes de lo que ahora conocemos como visualización de información algunos los atribuyen a William Playfair con su libro "Atlas de comercio y política (1786)". William Playfair era un ingeniero y politólogo escocés con un gran interés en las representaciones gráficas de datos estadísticos, que entendía que una representación gráfica era la mejor manera de comunicar la información contenida en un conjunto de estadísticas a personas que no estaban familiarizadas con el análisis estadístico. En dicho libro, Playfair intentó comunicar todas las estadísticas que se tenían sobre comercio en la Gran Bretaña del siglo XVIII. Como resultado de este esfuerzo, diseñó tipos de gráfica que seguimos utilizando a la fecha. Por ejemplo, este es el primer documento donde se utilizan las gráficas de barras. En esta época el gobierno de Gran Bretaña estaba comprometido con el registro y análisis estadístico de todos los datos posibles sobre comercio que se estaban generando, sin embargo, esta información era útil sólo para unos cuantos analistas especializados y no para el público en general, en particular, no la podían procesar personas involucradas en la toma de decisiones para los . Playfair dice al respecto: "Los personajes de muy alto rango no tienen tiempo para entrar en detalles sobre la información que se está generando y que requieren de métodos que les permitan tener acceso a la información de manera concisa y clara para estar al tanto" [6] (traducción libre). Actualmente sigue siendo fundamental la comunicación de los análisis de datos para la toma de decisiones, un ejemplo de esto es la popularidad de los Dashboards en el área de Businnes Intelligence, donde el principio es el mismo –sólo que en lugar de llamarlos personalidades de alto rango, se les llama ejecutivos corporativos–.

Después del trabajo de Playfair, tenemos un par de ejemplos emblemáticos de visualizaciones que permitieron descubrir o comunicar una idea. En 1854 ante una epidemia de cólera en el centro de Londres, John Snow, médico británico, visualizó los casos de muerte por cólera en un mapa del centro de Londres 2.1. Cada caso aparece como una rayita horizontal en el domicilio donde ocurrió, el mapa es un mapa detallado del centro de la ciudad, en particular, se pueden apreciar las bombas de agua que en esa época surtían a los ciudadanos del líquido. Esta visualización muestra claramente cómo los casos se apilan alrededor de una bomba específica, lo cual llevo a Snow a la conclusión de que la epidemia era provocada por agua contaminada proveniente de la bomba en cuestión. Cabe mencionar que en esa época no se sabían las causas del cólera y en realidad se pensaba que era transmitida por vía aérea. Por dicha razón la hipótesis de Snow de que se transmitía a través del agua encontró gran resistencia dentro de la comunidad médica y científica de la época. Al final, sin embargo, las pruebas fueron contundentes y se aceptó la causa.

Otro ejemplo es el de Florence Nightingale, enfermera británica conocida como “La dama de la lámpara” y cuyas ideas llevaron a la profesionalización de la enfermería. Dentro de las ideas de Nightingale estaba la necesidad de tener buenas condiciones de higiene para el cuidado de enfermos. Después de trabajar en los campamentos del Ejército del Este durante la guerra de Crimea, estas ideas la llevaron a emprender una batalla por mejorar las condiciones de los enfermos y heridos en los campamentos.

Nightingale además de ser enfermera era estadística, y entendía la conveniencia de la representación visual de datos para su comunicación. Como apoyo a sus argumentos sobre la necesidad de mejores condiciones de higiene en los hospitales del ejército inventó la gráfica de área polar, la cual usó para mostrar como se repartían las muertes de acuerdo con tres categorías: muertes por heridas de guerra, muertes por causas prevenibles y muertes por otras causas [Fig. 2.2]. En su análisis resultaba que las causas de muerte por complicaciones derivadas de la falta de higiene eran mayores a las muertes por heridas de guerra, y esto se podía ver claramente en su diagrama de área polar. Gracias a

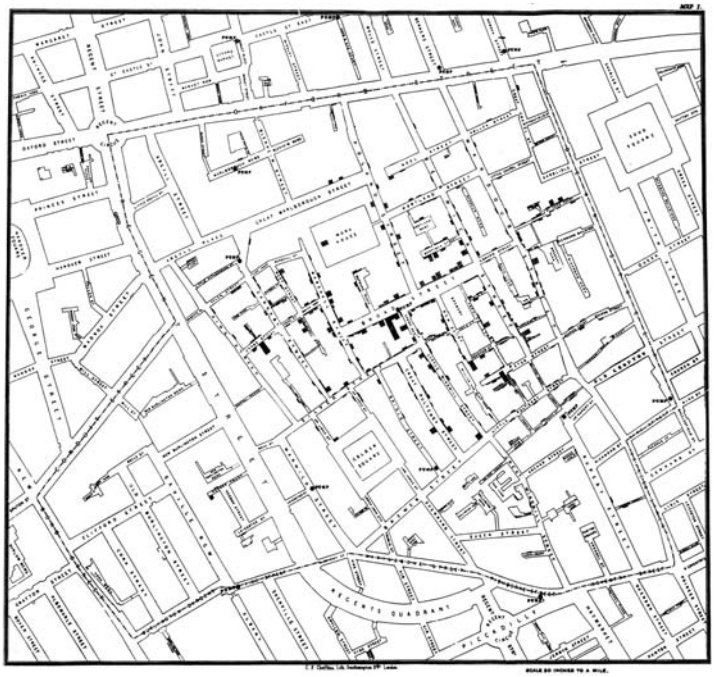


Figura 2.1: Mapa del centro de Londres que muestra los cúmulos de casos de cólera durante la epidemia ocurrida en 1854. Cada caso se muestra como una línea negra en el domicilio donde ocurrió. Se puede observar cómo los casos se apilan alrededor de un hidrante en Broad Street y Little Windmill Street. John Snow, 1854. Fuente: http://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg

esta representación gráfica, y a su manejo en general de las gráficas estadísticas como medio de comunicación, actualmente se considera que Nightingale además de ser fundadora de la enfermería moderna, es pionera en la presentación visual de información y gráficas estadísticas [59].

Una gráfica que muestra cómo un buen diseño puede comunicar una gran cantidad de información fue hecha por Charles Minard en 1869, ingeniero civil francés, que en una gráfica representa las pérdidas humanas durante la campaña de Napoleón para conquistar Rusia [Fig. 2.3], en qué partes del trayecto se dieron y qué eventos fueron los de mayores consecuencias. Esta gráfica, considerada por algunos como la mejor gráfica en la historia de las gráficas informativas, presenta el avance y la retirada de las tropas, combinando información sobre las temperaturas que sufrió el ejército

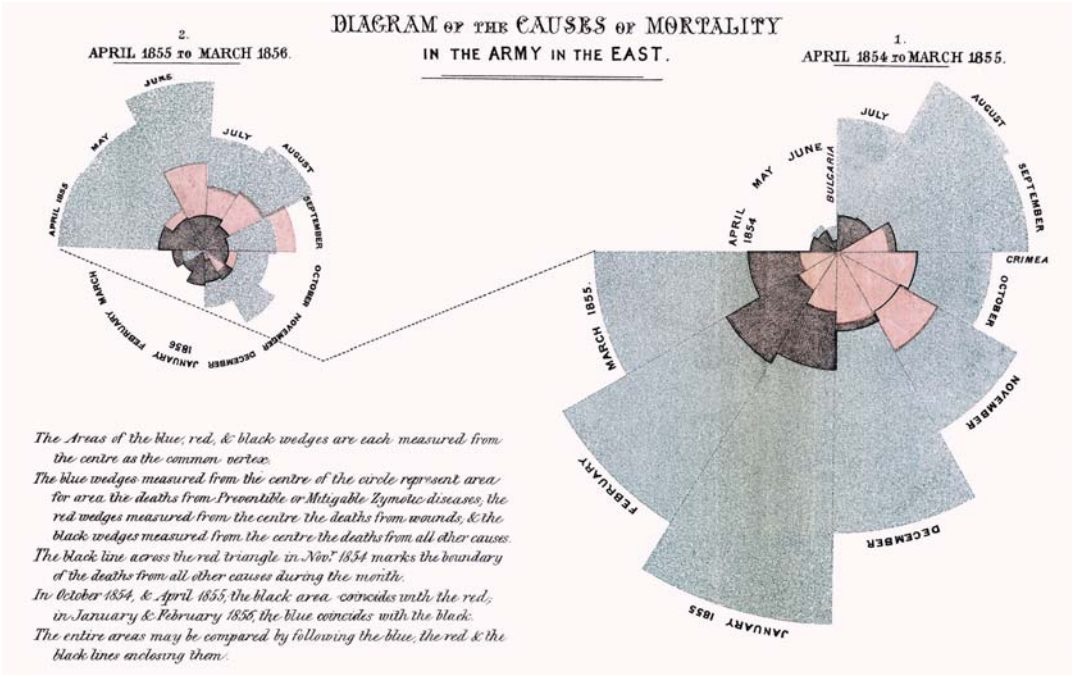


Figura 2.2: Diagrama de área polar que muestra estadísticas de las bajas en el Ejército del Este durante la guerra de Crimea entre abril de 1884 y abril de 1856. Cada color representa en área medida desde el centro, el número de muertos en un mes. La secciones azules representan muertes prevenibles, las secciones en rosa representan muertes por heridas de guerra, y las secciones en negro muertes por otras causas. Florence Nightingale, 1858. Fuente: http://es.wikipedia.org/wiki/Archivo:Nightingale-mortality.jpg

francés, representando también cómo las tropas se fueron adelgazando a lo largo del trayecto, donde quedan plasmadas en simples dibujos estadísticos grandes tragedias. Al final uno puede ver la brutal comparación entre el tamaño de las tropas que entraron al territorio ruso y el tamaño de las que salieron, representado por el grosor de las barras café (tropas en dirección a Moscú) y negra (tropas en retirada), donde un milímetro (en el diseño original) representa 6000 hombres. Está gráfica nos da información sobre las temperaturas en la retirada y las fechas en que se dieron. Asimismo nos da el tamaño de las tropas, su localización y dirección, además marca algunos puntos geográficos relevantes. En total nos permite hacer una análisis multivariado representando 6 variables en la misma gráfica [54].

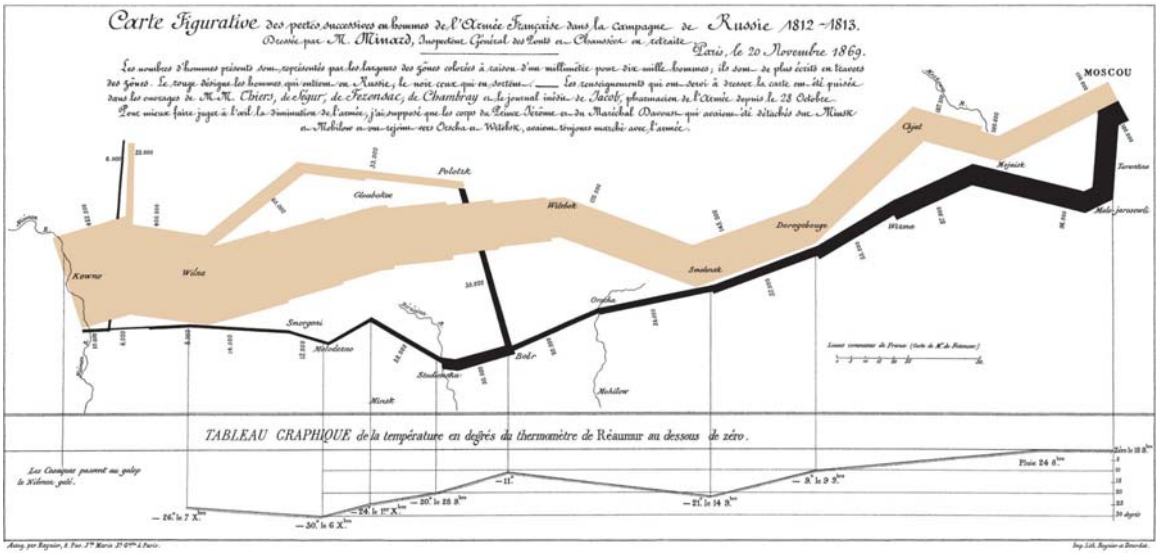


Figura 2.3: Carta figurativa de las pérdidas en hombres de la Armada francesa durante la campaña de Rusia 1812-1813. La barra café representa las tropas del ejército francés en dirección a Moscú, mientras que la negra representa las tropas en retirada. En ambas barras un milímetro de ancho (en el original de 55.88 x 50.8 cm) representaba 6000 hombres. Charles Minard, 1869. Fuente: http://en.wikipedia.org/wiki/File:Minard.png

Los párrafos anteriores son un breve recuento de algunos de los pioneros en el diseño de visualizaciones estadísticas, y que ejemplifican la importancia de traducir los datos en visualizaciones. Estos desarrollos fueron dando las bases de lo que es hoy la visualización de información. Sin embargo, después del auge de la visualización de datos en el siglo XIX, esta cayó en el olvido durante la primer mitad del siglo XX, periodo en el cual se vio más bien el surgimiento del interés en las ciencias sociales por modelos cuantitativos formales de carácter estadístico, y se consideraban las visualizaciones como 'sólo imágenes' [21].

El resurgimiento de la visualización vino en la segunda mitad del siglo XX. Uno de los hitos que marcaron este regreso fue publicación del libro del cartógrafo y teórico francés Jacques Bertin, "Semiótica de las gráficas (1967)". En este libro, Bertin formaliza el proceso de traslación de la información como pensamiento, al espacio visual como gráfica. Define la estructura básica de la

información y de las gráficas como sistema simbólico y a partir de eso fundamenta los pasos del análisis de información. La teoría presentada en este libro sorprende por su claridad y su visión sobre el cambio fundamental que estaba por darse con el uso de las computadoras para la visualización de datos. Este cambio consistía en la transición de la gráfica como imagen estática a la imagen dinámica, lo que permite que las gráficas pasen de ser un producto final que simplifica y comunica, a una herramienta de investigación, cuyo uso comienza desde el inicio del proceso de descubrimiento, y termina en la comunicación del conocimiento adquirido.

En el prefacio de la primera edición inglesa (1983), Bertin dice:

> "Ahora sabemos que entender quiere decir simplificar, reducir la vasta cantidad de datos al pequeño número de categorías de información que somos capaces de manejar al lidiar con un problema dado. Investigación en psicología experimental sugiere que los seres humanos podemos manejar tres y como máximo siete. El procesamiento de información involucra encontrar los métodos más aceptables para obtener esta simplificación indispensable"

Refiriéndose a los avances en la automatización del análisis de datos:

> "Con los avances en computación están al alcance toda clase de comparaciones, que anteriormente requerían ser hechas a mano mediante una serie de simplificaciones por el investigador. Sin embargo, simplemente alimentando datos a una computadora no constituye en sí mismo investigación científica. Hemos aprendido que las simplificaciones más importantes no son las automatizadas, sino las que preceden y suceden al análisis automático" (traducción libre).

El poder de cómputo con el que contamos nos ofrece medios muy poderosos para tratar los datos, y estos medios van en incremento en cantidad y en refinamiento. Entonces, por un lado tenemos una

gran cantidad de datos y por otro una gran cantidad de algoritmos para procesar esos datos. Esto presenta un problema: para responder a una pregunta no sólo debemos decidir qué datos son los más adecuados, y qué algoritmos son los más adecuados. Todo esto sin que nuestras capacidades de percepción hayan cambiado en lo más mínimo. En este sentido necesitamos herramientas que nos ayuden a dirigir nuestra búsqueda por el proceso óptimo de análisis.

Bertin, también define la relación entre las características de la información (cuántos componentes la conforman) y el tipo de preguntas que se pueden hacer a partir de esa información.

Pocos años después, en 1972, se creó el primer programa de gráficas estadísticas interactivas PRIM-9 [20], diseñado por John Tukey durante una estancia en el Computation Research Group del Stanford Linear Accelerator Center. PRIM-9 permitía al usuario proyectar en la pantalla los datos por pares de coordenadas agregando una tercera coordenada mediante rotaciones interactivas que permitían al usuario percibirla en 3era dimensión. Los conceptos básicos de interacción en PRIM-9 Tukey los nombró: “picturing”, “rotating”, “isolating”, “masking”. De los cuatro, el único que no se arraigó es masking. Tukey tenía muchas otras ideas sobre herramientas interactivas para el análisis de datos multivariados, algunas de las cuales, al parecer, sólo él comprendía y podía utilizar. Fue un visionario que comprendió el gran potencial para el análisis de datos que se guardaba en las computadoras.

Tukey, a su vez, estaba poco interesado en la formalización de herramientas. Su principal objetivo era la exploración de los datos y que esta exploración resultara intuitiva, cuando detectó que los usuarios tardaban en sus análisis al usar PRIM-9 y se volvía una tarea tediosa desarrolló una técnica a la que llamó “búsqueda de proyección” (projection pursuit), la cual servía para guiar al usuario en la búsqueda de proyecciones útiles de los datos. Esta idea fue desarrollada más adelante para desarrollar otros modelos como projection pursuit regression, un método con características similares a las de las redes neuronales. Lo interesante de esto es que Tukey no estaba pensando en un método de reducción de dimensiones que busca la proyección óptima. Su objetivo era proveer

de una herramienta para explorar el espacio de proyecciones y permitir al analista encontrar la que mejor servía a sus propósitos.

Otra idea innovadora eran las cognostics, proyecciones de los datos en dos dimensiones cuya finalidad era ser procesadas por la computadora para rankearlas de acuerdo con un índice de interés y presentarlas al analista para su examen final. La idea era que cuando se tienen muchas variables el número de proyecciones es demasiado para que sea analizado por un ser humano.

Estas ideas no sólo han influido el diseño de herramientas para análisis visual de datos hasta nuestras fechas. La idea fundamental que lo motivaba era la de proveer a los usuarios de metodologías para buscar estructuras interesantes en sus datos que los guiaran incluso en la búsqueda de nuevas estructuras. Él fue el principal promotor del análisis exploratorio de datos y su libro es un clásico con ideas aún vigentes.

2.1.3. Desarrollos recientes

El desarrollo de nuevas técnicas de visualización ha continuado. Un tipo de visualización de creación más reciente es el treemap [50], por Ben Shneiderman. Según relata Shneiderman, este tipo de visualización lo desarrolló gracias a su obsesión por saber cómo estaba usando su disco duro. Actualmente este tipo de visualización de datos jerárquicos se ha utilizado en diversas aplicaciones como la visualización de noticias, de datos financieros, y se sigue usando para visualizar en qué ocupamos nuestros gigas de disco duro.

En el análisis de datos, existen dos representaciones de los datos, la representación en la base de datos (cómo se guardaron los datos) y la representación mental (cómo entendemos los datos). La segunda es la que nos interesa, los datos se representan mentalmente de acuerdo con las preguntas que motivan el análisis y de acuerdo con las características de los datos. Por ejemplo, en un conjunto de datos demográficos sobre desempleo, uno puede preguntarse en qué zonas territoriales hay

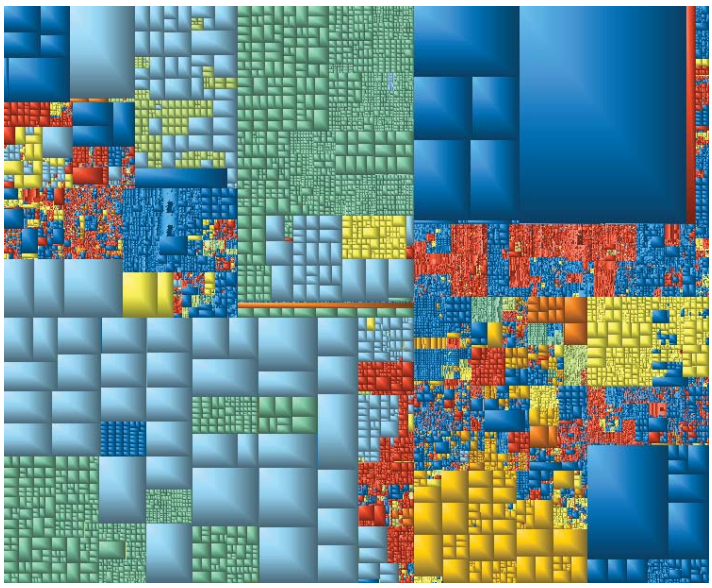


Figura 2.4: Visualización de treemap de un sistema de archivos usando la aplicación GrandPerspective. El espacio completo representa un directorio y los rectángulos que lo subdividen los subdirectorios o archivos contenidos en el directorio. El color de un rectángulo indica el tipo de archivo, y su área el porcentaje del total de memoria que ocupa el archivo.

mayor desempleo. En este caso se está pensando que las zonas son referidas por sus promedios de desempleo. Teniendo el mismo conjunto de datos uno puede preguntarse, ¿cuál es el promedio de desempleo en determinada zona? Esa pregunta implícitamente considera los datos de desempleo referidos por zona territorial. El objetivo del análisis visual es ayudar a comprender un fenómeno a partir de los datos disponibles, para esto la visualización debe estar diseñada pensando en la estructura de los datos conforme a las preguntas que se quieren responder. De esto se ocupa Bertin dedicando un capítulo completo a formalizar el concepto de "conjunto de datos datos". Esta formalización es actualizada por los autores de [2]. Para ellos, un conjunto de datos tiene dos tipos de componentes: los de referencia y los de características, donde los de referencia son, por lo general, de población (pensado como conjunto de objetos), de tiempo o de espacio.

2.2. Inferencia

De acuerdo con el diccionario de la Real Academia, inferir tiene las siguientes acepciones:

1. tr. Sacar una consecuencia o deducir algo de otra cosa. U. t. c. prnl.

2. tr. Llevar consigo, ocasionar, conducir a un resultado.

3. tr. Producir o causar ofensas, agravios, heridas, etc.

En nuestro caso nos referimos a la primera acepción y hay que distinguir la definición de "inferencia" en el lenguaje común de la definición de "inferencia" en estadística. La inferencia estadística es un intento por formalizar y cuantificar los procesos de inferencia que seguimos los seres humanos aprovechando nuestro conocimiento empírico. Inferencia quiere decir sacar conclusiones a partir de la información que se tiene a disposición. Hablando en términos de estadística tenemos que ser más precisos ¿qué se usa para sacar esas conclusiones? y ¿cómo se interpreta la información a disposición?

Actualmente, existen dos corrientes principales de pensamiento estadístico: la frecuentista y la bayesiana. La primera fue el estándar durante los 1900, mientras que la segunda, por largo tiempo desestimada, comenzó a recobrar importancia recientemente. Una razón para este resurgimiento es que gracias a las nuevas herramientas computacionales se pueden aplicar sus métodos efectivamente. Esto último aunado a que para algunos su forma de entender la probabilidad permite plantearse los problemas de inferencia de manera más natural. La corriente bayesiana entiende las preguntas en estadística como problemas que se derivan de la falta de información con respecto a un fenómeno, y por lo tanto de la incertidumbre de los modelos. Para la corriente frecuentista se plantean los fenomenos como resultado de un proceso aleatorio, para los bayesianos esta visión complica el razonamiento acerca de los fenómenos. Por ejemplo, si queremos evaluar si una moneda está cargada, para el bayesiano este es un problema de falta de información, pero no quiere decir

que el que la moneda este cargada sea una variable aleatoria, es decir, la moneda está o no está cargada. En el caso del pensamiento frecuentista se plantea el que la moneda esté cargada como una variable aleatoria. En el libro [52] los autores ponen como ejemplo los cálculos de Laplace para estimar la masa de Saturno, en donde lo que en realidad hizo fue calcular la función de densidad de probabilidad para la masa. Los autores argumentan que en la visión frecuentista se plantearía la masa como una variable aleatoria, lo cual es poco natural y sólo complica el razonamiento de inferencia porque la masa de Saturno no es una variable aleatoria, es un valor fijo. Lo que pasa es que no sabemos cuál es, pero lo que podemos hacer es acercarnos a su valor utilizando la información que tengamos a la mano.

2.2.1. Experimentos controlados y estudios de observaciones

En el análisis de un fenómeno se pueden tener distintas condiciones en cuanto a los recursos disponibles para el desarrollo de una teoría. En una situación los datos son generados por el investigador en un ambiente controlado. En este caso el investigador realiza una serie de experimentos donde genera los datos que necesita y compara los resultados de acuerdo con variaciones inducidas por él mismo. Por ejemplo, si se quieren medir los efectos de un fertilizante uno podría hacer el estudio con una población de plantas, abonar la mitad de estas e ir midiendo su desarrollo a lo largo de un periodo de tiempo y comparar los resultados entre las que fueron abonadas con el fertilizante y las que no. A este tipo de situaciones se conocen como experimentos controlados y el diseño de experimentos es una rama de la estadística que se encarga de determinar las mejores condiciones para implementarlos.

Otra situación común es cuando no se tiene control sobre la forma en que se generan los datos para el análisis. Las razones por las que el investigador se puede encontrar en esta situación son diversas. Por ejemplo, para establecer los efectos secundarios que tiene un medicamento en niños no sería

ético darle el medicamento a niños para ver qué reacciones presentan. Otra situación sería que simplemente no se puedan inducir las condiciones para generar los datos. Esto es común en estudios sociodemográficos. Un ejemplo sería un estudio sobre los factores que inciden en la intención de voto, no hay forma de manipular factores como la economía. En dicho caso el sociólogo interesado tendrá que hacer encuestas y buscar que factores parecen estar correlacionados con la intención de voto, es importante notar que no podrá comprobar causalidad, dado que no tiene control sobre las variaciones de los factores que detecte como significativos. Otro problema es que muy probablemente el sociólogo no tenga los recursos para preguntar a todos los ciudadanos, por lo que tendría que hacer un muestreo de la población, esto agrega un nivel más de incertidumbre, y su control será únicamente sobre la forma en que se hace el muestreo. En el peor de los casos, uno tiene datos que no fueron recolectados con el propósito del análisis que se quiere hacer, y son los únicos que puede conseguir. En este caso la necesidad de hacer análisis exploratorio es indispensable.

2.2.2. Prueba de hipótesis

En el análisis de datos, además de construir hipótesis, también se requiere validarlas, o en su defecto tener algún parámetro que nos indique su grado de confiabilidad. Es decir, dada una hipótesis nos gustaría saber si la evidencia apoya esta hipótesis, o si por el contrario, la contradice. En estadística existe una serie de técnicas que permiten realizar este tipo de evaluaciones, conocidas como pruebas de hipótesis.

En análisis de datos, las hipótesis son proposiciones acerca de la naturaleza de algún fenómeno, y según la perspectiva desde la cual se generen pueden estar definidas de maneras muy distintas. En estadística clásica una hipótesis toma la forma de la estimación de un parámetro, por ejemplo, el valor esperado de una variable aleatoria o la probabilidad de cierto resultado pueden ser formuladas como hipótesis. Dada una moneda, podemos empezar con la hipótesis de que la moneda

no está cargada y que "el número esperado de águilas en diez volados es cinco". Esta hipótesis tendría que ser corroborada lanzando volados y cuantificando si el resultado soporta o rechaza esta hipótesis (o ninguna de las anteriores). Para probar esta hipótesis, la técnica de libro es postular una hipótesis alternativa y probar cuál tiene mayor sustento según los datos de muestreo. Por ejemplo, una hipótesis alternativa es, para la misma moneda, "el número esperado de águilas en diez volados es tres".

Supongamos que en un juego de dados un jugador sospecha que un dado está cargado y que favorece al cuatro. En general, a la hipótesis que se busca contradecir se le llama hipótesis nula, en este caso la hipótesis nula es que el dado no está cargado, lo que implica en términos de probabilidad que la probabilidad de que salga cuatro en un tiro es $P(X = 4) = \frac{1}{6}$. El jugador incrédulo entonces postula que en este dado $P(X = 4) > \frac{1}{6}$. Esta sería la hipótesis alternativa. Para decidir si es cierto que el dado está cargado, los jugadores lanzan el dado 12 veces. El resultado es que el cuatro sale en seis ocasiones, lo cual provoca la indignación del jugador y éste se retira, ¿tiene razón en indignarse? El número de cuatros esperado en 12 tiros es dos si el dado no está cargado, pero eso no quiere decir que cada vez que tiremos el dado 12 veces saldrá exactamente en dos ocasiones. Entonces la pregunta es, ¿cómo decidimos que el que salga cuatro veces es evidencia de que está cargado? La respuesta es calcular qué tan probable es que se dé ese resultado si el dado no está cargado. Además de la hipótesis nula y la hipótesis alternativa, se requiere una región de rechazo. Esta última se refiere al intervalo de valores con los que se rechaza la hipótesis nula.

En el caso de los jugadores de dados se podría definir que si el número de cuatros en 12 tiros es mayor que cinco entonces el dado está cargado –se rechaza la hipótesis nula. Dado el criterio de rechazo otra pregunta razonable es, ¿cuál es la probabilidad de que nos equivoquemos si usamos esta región de rechazo? Se pueden dar dos tipos de error, el error tipo I, sería descartar la hipótesis nula cuando es verdadera. En nuestro ejemplo esto sería que salgan cinco cuatros o más y el dado no esté cargado. El otro tipo, error tipo II, es que aceptemos la hipótesis nula cuando en realidad es

falsa –el dado sí está cargado, pero salieron menos de cinco cuatros–. En resumen, los errores se refieren a la probabilidad de tener un falso negativo y un falso positivo.

La probabilidad de error tipo I se le denota como α y en este ejemplo se calcularía de la siguiente manera

$$\alpha = \sum_{i=6}^{12} \binom{12}{i} \left(\frac{1}{6}\right)^{i} \left(\frac{5}{6}\right)^{12-i} = 0.00792504 \tag{2.1}$$

Esto es, α es la probabilidad de que en 12 tiros obtengamos seis o más cuatros si el dado no está cargado –la probabilidad de un cuatro es $\frac{1}{6}$–. El error tipo II, se denota por β. Para calcularlo supongamos que el dado está cargado y la probabilidad de que salga cuatro es $\frac{1}{4}$, entonces tendríamos que

$$\beta = \sum_{i=0}^{5} \binom{12}{i} \left(\frac{3}{4}\right)^{i} \left(\frac{1}{4}\right)^{12-i} = 0.9455978 \tag{2.2}$$

nos da la probabilidad de que salgan menos de seis cuatros cuando la probabilidad de que salga un cuatro es de $\frac{1}{4}$, la cual es muy alta, es decir lo más probable es que no detectemos si el dado cargado a favor del cuatro.

En una prueba de hipótesis queremos minimizar ambos tipos de error y para esto es clave la elección de la región de rechazo. Por desgracia, minimizar un tipo de error conlleva el incremento del otro. Para ver esto consideremos los casos extremos, la región de rechazo es vacía, o sea, siempre rechazamos la hipótesis nula con lo cual $\alpha = 0$, pero se maximiza β. En el otro extremo, la región de rechazo es el total de posibles resultados, lo que implica que siempre aceptamos la hipótesis nula, por lo tanto $\beta = 0$, y maximizamos α. El objetivo es encontrar un punto de equilibrio con el que se logre llevar ambos tipos de error a un nivel aceptable. Claramente, si se tiene la posibilidad de aumentar a su vez el número de observaciones esto disminuye la probabilidad de ambos tipos de error.

En el ejemplo anterior nuestras hipótesis se hicieron alrededor de la probabilidad de obtener un resultado en el dado. En general las pruebas de hipótesis se construyen alrededor de distintas estadísticas. Por ejemplo, el valor esperado de una variable aleatoria o la desviación estándar. En estos casos se postula como hipótesis nula un valor para la estadística. En el caso anterior, la hipótesis nula se puede plantear usando el valor esperado del número de cuatros en doce tiros, $\mu_N = 2$, contra el valor esperado estimado a partir de los resultados del experimento. Supongamos que se obtuvieron seis cuatros, entonces $\mu_A = 6$. En este caso la comparación directa de los dos valores no es adecuada para definir si la moneda está cargada, un factor importante del que carece tal comparación es que no se considera el número de eventos, el cual es determinante en el grado de certeza que tenemos sobre la conclusión. En la siguiente sección discutimos una estadística de diagnostico para prueba de hipótesis, en la cual sí se considera el número de eventos.

Este tipo de pruebas de hipótesis tienen algunos problemas, entre ellos, que el umbral que se plantee para rechazar o aceptar una hipótesis no deja de ser arbitrario, es decir, ¿por qué se rechaza si está fuera del 95 %? Si está entre el 94.9 % y el 95 % se acepta, pero si está entre el 95 % y el 95.01 %, ¿se rechaza? Por qué no rechazar sólo si está fuera del 99.9 %. Esta es una de las críticas que uno se encuentra en la literatura. Por esta razón es importante que el analista entienda la estructura de las relaciones y tome decisiones informadas, en lugar de aplicar reglas de libro de texto indiscriminadamente.

En general, dado que tenemos información limitada con respecto a los fenómeno que estudiamos, queremos ver si una hipótesis puede ser mejorada en cuanto a su capacidad para modelar el sistema. Esa es la idea de fondo en la prueba de hipótesis, donde se busca contrastar lo que uno cree con respecto a una alternativa específica.

En particular en el análisis de datos una tarea común es determinar si las variables del sistema se afectan entre sí, o si nos dan algo de información unas sobre las otras. En términos estadísticos buscamos ver si hay algún tipo de correlación entre las variables. Un método para determinar esto

es considerar las dos hipótesis, es decir, que las variables no se afectan y que las variables sí se afectan y comparar cual se sustenta más con nuestros datos, esto a grandes rasgos es lo que hace el método que explicamos a continuación, y que será nuestra herramienta de diagnóstico de relaciones a lo largo de la tesis.

2.2.3. Épsilon

Sean X y Y dos variables aleatorias para las que queremos saber si están correlacionadas o si una depende de la otra. Si X y Y son variables booleanas (variables que toman uno de dos valores comúnmente asociados con valores de "verdad"), el objetivo es saber si el valor que se obtenga en una observación de una de las variables nos da información sobre el valor que podría tener la otra variable. Un primer intento para medir si existe alguna dependencia es considerar las probabilidades $P(X)$ y $P(X \mid Y)$ –donde $P(X)$ es la probabilidad de que X sea "verdad" y $P(X \mid Y)$ es la probabilidad de que X sea "verdad" dado que Y es "verdad"–, y comparar estas cantidades. Esto nos daría una idea sobre si X es dependiente de Y. Si tenemos N tales que en N_X de los casos X fue "verdad" y en N_Y de los casos Y fue "verdad", entonces podemos estimar la probabilidad $P(X)$ de que X sea "verdad" con:

$$\hat{P}(X) = \frac{N_X}{N} \tag{2.3}$$

Si además tenemos que en N_{XY} evento tanto X como Y fueron "verdad" entonces podemos estimar la probabilidad $P(X \mid Y)$ de que X sea "verdad" cuando Y es "verdad" con:

$$\hat{P}(X \mid Y) = \frac{N_{XY}}{N_Y} \tag{2.4}$$

Si consideramos la diferencia $P(X \mid Y) - P(X)$ o el cociente $\frac{P(X|Y)}{P(X)}$ estas cantidades nos indican si hay más (o menos) probabilidad de que X sea "verdad" cuando Y es "verdad", lo cual podría

indicar que existe cierta dependencia. Pero, esto no es condición suficiente, ya que no sabemos si la diferencia fue generada por efectos del azar. Por ejemplo, si X representa obtener águila al echar un volado y Y representa que estén jugando los Pumas, y echo una serie de 10 volados cuando están jugando y otra serie de 10 cuando no están jugando, lo más probable es que no salga el mismo número de águilas en cada experimento, con lo cual $P(X)$ y $P(X \mid Y)$ serían distintas y por lo tanto podría concluir, erróneamente, que el que jueguen los Pumas da información sobre el número de águilas que uno obtendría en una serie de volados en ese momento.

Lo que faltó en el razonamiento anterior es que no estamos considerando que la diferencia es también una variable aleatoria. Lo que necesitamos es una forma de medir si la diferencia es suficientemente grande como para descartar que proviene de la naturaleza aleatoria de N_X, N_{XY} y N_Y. Lo primero que notamos es que X y Y son variables booleanas, y que el número de ocurrencias define un variable aleatoria con distribución binomial. Sea S_X el número de ocasiones en que una variable booleana con probabilidad $p_x = P(X)$ es "verdad" en N_Y eventos, entonces S_X es una variable binomial y tenemos que

$$\mu_{S_X} = N_Y p_x \tag{2.5a}$$

$$\sigma_{S_X} = \sqrt{N_Y p_x (1 - p_x)} \tag{2.5b}$$

Análogamente, si $S_{X|Y}$ es el número de ocasiones que en que una variable booleana con probabilidad de ser "verdad" $p_{x|y}$ sería "verdad" en N_Y eventos, entonces

$$\mu_{S_{X|Y}} = N_Y p_{x|y} \tag{2.6}$$

Una forma de medir si la probabilidad de X se ve afectada por la presencia de Y es comparar el

valor esperado de S_X con el de $S_{X|Y}$ como una prueba de hipótesis, es decir por un lado la hipótesis nula de que X es independiente de Y es representada por el valor esperado de la variable S_X y la hipótesis alternativa X depende de Y es representada por el valor esperado de $S_{X|Y}$. Si utilizamos la desviación estándar de S_X para estandarizar la diferencia entre los valores esperados obtenemos la estadística de prueba de hipótesis ϵ

$$\epsilon_{x,y} = \frac{\mu_{S_{X|Y}} - \mu_{S_X}}{\sigma_{S_X}} \tag{2.7a}$$

$$= \frac{N_Y(\hat{P}(X|Y) - \hat{P}(X))}{\sqrt{N_Y \hat{P}(X)(1 - \hat{P}(X))}} \tag{2.7b}$$

Una propiedad de la distribuciones binomiales es que si el número de eventos es suficiente pueden ser aproximadas por una distribución normal (este resultado es conocido como el Teorema del límite central) en cuyo caso nuestra región de rechazo de la hipótesis nula es fácil de determinar. En una distribución normal con valor medio μ y desviación estándar σ los posibles resultados se distribuyen de tal forma que el 95 % cae dentro del intervalo $[\mu - 2\sigma, \mu + 2\sigma]$, por otro lado, una distribución normal es simétrica, es decir, sólo 2.5 % de los posibles resultados son mayores que $\mu+2\sigma$ y 2.5 % menores que $\mu-2\sigma$. Por lo tanto, volviendo a ϵ, si existe un número de observaciones suficiente y $\epsilon_{x,y} > 2$ la probabilidad de que por azar se haya dado la diferencia de valores es de .025, es decir, la certeza de que existe algún tipo de dependencia o interacción es alta y por lo tanto se descartaría la hipótesis nula de que X es independiente de Y.

En la construcción de ϵ sólo se consideró que X y Y son variables booleanas, por lo que nada impide que cada una de estas sea a su vez una composición de variables booleanas. Si tenemos n variables booleanas $X_1, \ldots, X_n$, podemos considerar variables construidas a partir de funciones booleanas $X = F_x(X_1, \ldots, X_n)$ y $Y = F_y(X_1, ..., X_n)$, por ejemplo se pueden definir $Y = X_1X_3 \vee X_6$ y $X = X_4X_5$ y estimar $P(F_x(X_1, \ldots, X_n) \mid F_y(X_1, ..., X_n))$.

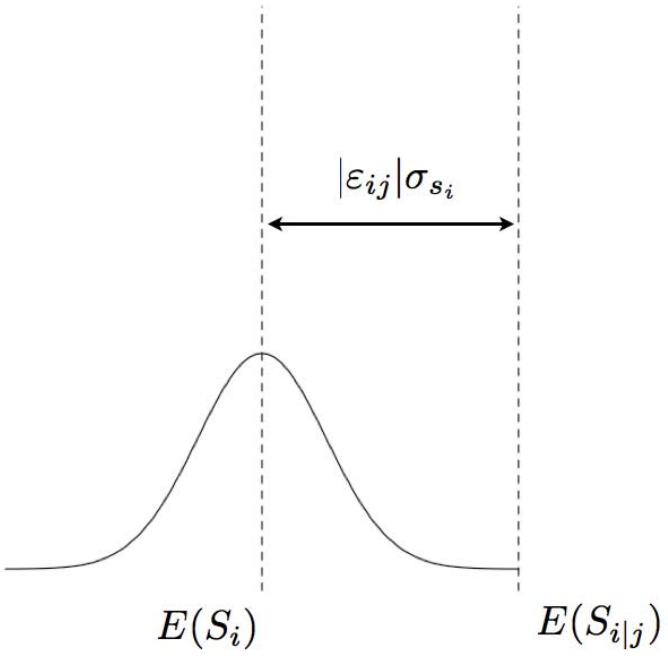


Figura 2.5: $\epsilon_{x,y}$ es la distancia entre μ_{S_X} y $\mu_{S_{X|Y}}$ en desviaciones estándar de S_X

Los usos de una estadística de diagnostico como ϵ son varios. El uso más directo es para aceptar o descartar la hipótesis nula de independencia de una variable booleana con respecto a otra. Pero hay más, por un lado el signo de ϵ nos indica si la relación es de 'atracción' o 'repulsión', positivo o negativo respectivamente, por otro la magnitud nos permite diagnosticar entre un conjunto de variables cuales son las que tienen mayor potencial como variables predictivas con respecto a nuestra variable o clase de interés. En este caso se puede hacer una selección de las variables con mayor potencial de predicción de acuerdo con la magnitud de ϵ, lo que se conoce como selección de características 2.3.

En el capítulo 3 se muestra como se puede construir y aplicar ϵ en el contexto de minería de datos espaciales, en el capitulo 4 se presenta la construcción de redes a partir de ϵ como técnica de exploración de datos y posteriormente se presenta un estudio en el que se combinan ambas cosas.

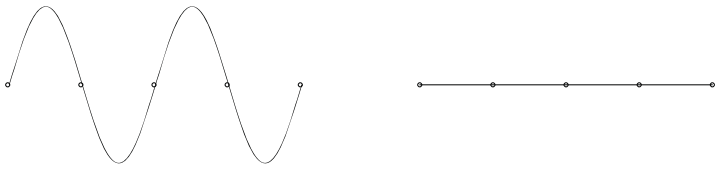
2.3. Reducción de dimensiones

Un problema importante en el análisis de datos es el número de variables involucradas, aunque por un lado entre más información tengamos a nuestra disposición es mejor, esto también implica un incremento en la complejidad del análisis. La complejidad proviene de que el tamaño del espacio de búsqueda crece rápidamente y el número de datos necesarios para establecer significatividad estadística crece exponencialmente conforme aumenta el número de dimensiones. Una parte fundamental consiste en determinar que variables dan poca información ya sea porque no están relacionadas realmente con el fenómeno o porque sólo repiten la información que dan otras variables. Este problema atormenta tanto a los analistas que ha sido bautizado como la "maldición de la dimensionalidad".

2.3.1. Sesgo vs Varianza

En problemas de modelación buscamos que el modelo describa lo mejor posible el comportamiento de las variables del sistema. Pero, ¿qué quiere decir esto? Por un lado podemos tener un modelo demasiado simple (sesgo), que no logra capturar la complejidad del comportamiento de las variables. Por otro lado podemos tener uno demasiado complejo (varianza), sobreajustado a los datos, que pierde la capacidad de explicar nuevos datos o de predecir. Encontrar este balance no es trivial. Un principio que se utiliza como guía para discernir entre dos modelos que capturan al mismo detalle los datos del sistema es elegir el más sencillo, este criterio se conoce como la 'Ley de parsimonia' o 'la navaja de Ockham', que sugiere que entre dos teorías con la misma capacidad explicativa es más probable que la más simple sea la verdadera. Aunque este criterio en su momento se consideró un principio o ley, queda claro que no forzosamente es cierto, no hay nada que nos asegure que las teorías más simples son las más probables. Sin embargo, para efectos prácticos, si tenemos dos modelos de los cuales no sabemos cuál es el más apegado a la realidad, nos conviene utilizar el más

simple. En el caso de la construcción de modelos es común que el que un modelo sea más simple este vinculado con la varianza asociada a este, con lo que si tenemos dos modelos que generan los datos, pero uno presenta más varianza una regla general es utilizar el modelo con menor varianza, esto aunque el modelo más simple no reproduzca exactamente los datos originales.



(a) (b)

Figura 2.6: Ejemplo de modelos que aproximan un mismo conjunto de datos, y donde la elección clara es el modelo lineal.

El problema es que no es fácil determinar cuando un modelo es suficientemente bueno para descartar uno más complejo que se ajusta mejor a los datos. Esta tarea requiere de la exploración de los datos para obtener la mayor información sobre los factores que pueden estar en juego en el comportamiento de los datos y desarrollar intuición sobre el tipo de modelo más adecuado para describirlos.

Reducir la dimensión de un conjunto de datos multivariado es importante tanto para desarrollar modelos predictivos como para descriptivos. A continuación se exponen algunos de los métodos de reducción de dimensiones más representativos en el área de análisis exploratorio. Los cuales usamos como ejemplo porque además muestran el tipo de decisiones que se tienen que tomar para intentar simplificar un conjunto de datos para llevarlo a una escala comprensible, y los sacrificios que implican las decisiones. PCA obtiene una proyección ortogonal en m dimensiones, de tal forma que la proyección captura la mayor parte de la varianza en los datos; MDS por otro lado tiene como objetivo bajar de dimensión pero manteniendo lo más posible distancias entre los puntos, es

decir, intenta preservar la geometría; finalmente, SOM, proyecta los datos a una dimensión menor buscando preservar topología.

2.3.2. Análisis de componentes principales (PCA)

El análisis de componentes principales es una técnica popular para el análisis exploratorio de datos multivariados, su objetivo es la reducción de dimensiones de conjuntos de datos. Se utilizan tanto para el reconocimiento de patrones, como para la exploración de datos y su visualización.

Como es el caso en los métodos de reducción de dimensiones, el objetivo que persigue el análisis de componentes principales es proyectar el conjunto de datos multivariados a un espacio de menor dimensión perdiendo la menor cantidad de información posible. Ahora bien, ¿qué quiere decir perder información? Este concepto varía de un método a otro y en general se ve reflejado en alguna función objetivo que se busca minimizar, en el caso del PCA se considera que la información está contenida en la varianza de los datos. Por ejemplo, si contamos con un conjunto de datos con dos variables y una es constante, para efectos de nuestro análisis la constante no nos da información, lo que nos interesa es analizar los cambios en la que no lo es.

Por lo tanto, para conseguir su objetivo, el PCA, busca el conjunto de direcciones en el espacio de datos en las que se da la mayor variabilidad [29, 65], a estas direcciones se les llama componentes principales, y proyectando los datos sobre estas direcciones se espera minimizar la pérdida de información, en el sentido de que se captura la mayor parte de la varianza.

En general, en PCA se obtienen n vectores, ortogonales entre ellos, que conforman un cambio de coordenadas en el espacio de datos. Si se ordenan estos vectores en orden decreciente, de acuerdo con la magnitud de la varianza de los datos sobre cada vector y se eligen los primeros k, con ellos se define una proyección en k coordenadas que capturan la mayor varianza.

Una propiedad de la varianza total de un conjunto de puntos es que esta es invariante ante cambios

de coordenadas, mientras sean sistemas de coordenadas ortogonales. Por lo tanto, si uno calcula la varianza total de los datos, uno puede calcular cuánto de esa varianza se perdió en la proyección, y si no se tiene una restricción para el número de dimensiones de la proyección uno podría elegir las k direcciones, $k < n$, tales que el porcentaje de la varianza total en la proyección sea mayor que un mínimo requerido, por ejemplo, tales que capturen el 90 % de la varianza. Sin embargo, para efectos de visualización, la dimensión por lo general esta restringida a dos o tres y en este caso la reducción de dimensiones puede ser demasiado grande.

El PCA funciona de la siguiente manera: Si se tiene un conjunto de n datos de dimensión p, estos se pueden ver como una matriz X de $n \times p$, donde cada renglón corresponde a un vector de datos y cada columna a una variable. La varianza a lo largo de una dirección a se define como

$$\sigma_a^2 = (Xa)^T(Xa) = a^T(X^TX)a \tag{2.8}$$

donde X^TX es la matriz de covarianza del conjunto de datos X, el objetivo para encontrar la dirección de máxima variabilidad es maximizar σ_a^2. Para que la solución esté bien definida se agrega la restricción de que a esté normalizado, de lo cual se deriva [29] que la solución está dada por

$$(X^TX - \lambda I)a = 0 \tag{2.9}$$

la forma clásica para definir los vectores propios de una matriz. En este caso la solución está dada por el vector propio de la matriz de covarianza X^TX con el valor propio más grande. Del mismo modo, los componentes principales subsecuentes corresponden a los vectores propios de la matriz de covarianza en orden decreciente según sus valores propios.

En resumen, el proceso de PCA consiste en: obtener los vectores propios de la matriz de covarianza X^TX; ordenarlos decrecientemente con respecto a sus valores propios respectivos; elegir las primeras k direcciones, las cuales corresponden a las k direcciones que contienen mayor variabilidad. Por

lo general, si usamos PCA para encontrar una transformación que nos permita visualizar los datos entonces en general buscamos que k sea dos o tres.

El PCA involucra principalmente álgebra matricial, por lo que el proceso está bien entendido desde el punto de vista matemático, cosa que no sucede con el mapeo auto-organizado, el cual explicamos más adelante. Además sus soluciones son robustas y en general eficientes para conjuntos de puntos no demasiado grandes. Una propiedad atractiva es que la proyección en la primer coordenada tiene la mayor varianza, la segunda coordenada tiene la segunda varianza más grande y así sucesivamente.

Otra ventaja es que el análisis es no paramétrico. Es decir, no hay que ajustar parámetros basados en la experiencia y la solución está bien definida. Sin embargo, esto también es una limitación porque implica que no podemos utilizar conocimiento *a prioiri* sobre el tema para influir en el proceso [49].

La limitación más fuerte del PCA es que no captura relaciones no lineales más allá de las definidas por estadísticas de segundo orden [65] debido a que es una proyección lineal. Por ejemplo, esto implica que si la solución óptima se encuentra en 2D pero no es un plano sino una superficie en 3D, el PCA probablemente tendrá un desempeño pobre. Uno de los efectos conocidos por esta limitación es el efecto de la herradura de caballo [EXPLICAR], limitación que se hace más importante conforme crece el tamaño del conjunto de datos o su dimensión. En especial, cuando se usa como método de visualización de datos, puede dar como resultado una visualización de poco valor.

Para resolver la limitación de linealidad en el PCA se han propuesto los PCA no lineales, o análisis de curvas y superficies principales, esta extensión del PCA se encuentra en ocasiones asociada en la literatura con el SOM [30]. El problema es que hasta ahora no existe una solución única y el tema sigue abierto. Aún así, ha servido para motivar soluciones prácticas [63].

Otro problema del PCA es que no se escala bien en su forma tradicional para dimensiones grandes. Calcular los componentes principales de las ecuaciones de vectores propios se escala a grandes

rasgos como $O(np^2+p^3)$(np^2 para calcular la matriz de covarianza y p^3 para resolver las ecuaciones de valores propios para la matriz de $p \times p$) [29], es decir, la complejidad del cálculo se incrementa rápidamente conforme aumenta el número de variables, lo cual lo vuelve muy ineficiente para análisis de datos con muchas variables.

Finalmente, en lo que se refiere a la visualización, la proyección de los datos puede ser difícil de entender para el usuario, en especial si no está familiarizado con los conceptos de álgebra lineal detrás de este método [49].

2.3.3. Escalamiento multidimensional (MDS)

El escalamiento multidimensional es otro método tradicional para reducir el número de dimensiones de un conjunto de datos a una escala más manejable. El objetivo que persigue el MDS es encontrar una transformación, tal que las distancias entre los objetos originales se mantengan lo más posible entre sus proyecciones. Estos métodos, en general definen una proyección no lineal por lo que tienen mayor capacidad de representación que los PCA.

Existen varios modelos de MDS, los cuales varían según sus objetivos, el espacio geométrico al que se quieren proyectar los datos, o los algoritmos usados para encontrar una representación óptima. Sin embargo, lo que tienen en común es que todos buscan modelar las diferencias entre un conjunto de datos en un espacio geométrico de menor dimensión.

Una propiedad de los MDS es que, en lugar de usar el conjunto de datos, sólo necesitan las distancias o disimilitudes entre objetos. Un ejemplo, es el siguiente experimento. A cierto número de sujetos se les pidió que al escuchar una clave del sistema Morse identificaran la letra correspondiente, a partir de esto se generó una matriz de similitudes, donde la similitud entre dos letras representaba que tanto se habían confundido los sujetos entre estas. En el caso del MDS esta matriz es todo lo que se necesita conocer, es decir no se necesitan saber los datos originales que dieron pie

a la medida de similitud. Esta característica da mayor flexibilidad al MDS ya que esto permite que pueda ser usado con datos no numéricos

En general, si tenemos un conjunto de objetos y para cada pareja de objetos (i, j) tenemos una medida de disimilitud D_{ij}, el modelo de MDS está dado por una función

$$f : D_{ij} \rightarrow d_{ij}(X) \tag{2.10}$$

donde X es una configuración de puntos en m dimensiones, y f es tal que las disimilitudes D_{ij} están bien aproximadas por las distancias d_{ij}, lo más común es que el mapeo de los objetos sea a un conjunto de puntos $x_1, ..., x_n \in R^m$ y las distancias d_{ij} estén dadas por $\|x_i - x_j\|$.

La versión de MDS más recurrida se conoce como MDS métrico y ha sustituido la versión clásica [9]. En el MDS métrico se busca minimizar una función de ajuste o *estrés* que es una medida de que tanto se diferencian las disimilitudes D_{ij} de las distancias d_{ij}. El caso más simple está dado por

$$S = (\sum_{i \neq j} (D_{ij} - \|x_i - x_j\|)^2))^{1/2} \tag{2.11}$$

Existen varias formas de definir la función de ajuste, otra forma conocida es la función de Sammon, donde se aplica una normalización intermedia para preservar buenas distribuciones locales y mantener la estructura global [64].

$$S = \frac{1}{\sum_{i<j} d_{ij}} \sum_{i<j} \frac{[d_{ij} - D_{ij}]^2}{d_{ij}} \tag{2.12}$$

En general, para funciones de *estrés* como las anteriores, se usan algoritmos de descenso de gradiente, esto puede generar problemas debido a mínimos locales y divergencia, además de que pueden ser computacionalmente intensivos [65].

Los MDS son métodos poderosos para revelar la estructura en los datos porque explícitamente

tratan de mantener las proporciones de las distancias entre parejas de puntos. Sin embargo, cuando se tiene un gran número de puntos, al igual que con PCA la estructura se vuelve poco evidente, además de que dado que involucran métodos altamente sofisticados es posible que se introduzcan efectos artificiales [29], otra crítica a esta clase de métodos es que no dan una función de proyección explícita por lo que para cualquier punto de datos nuevo el mapeo tiene que ser recalculdado [63].

2.3.4. Mapeos auto-organizados (SOM)

Otra técnica, de origen más reciente, utilizada en el análisis exploratorio de datos son los mapeos auto-organizados. Un mapeo auto-organizado, también conocido como red de Kohonen, es un tipo de red neuronal no supervisada que sirve para ordenar conjuntos de datos sobre una malla de dimensión menor a la de los datos–en general en una malla de 2 dimensiones–, de tal manera que la distribución sobre la malla representa las similitudes que existen entre los datos originales.

El resultado del algoritmo de aprendizaje del SOM es una red neuronal donde los nodos están organizados en un malla regular y la distancia entre nodos expresa el grado de similitud que hay entre los datos asociados a los nodos. El SOM se utiliza tanto como método de detección de cúmulos (clustering) como para visualización de datos, y al igual que el MDS produce por lo general mapeos no lineales [63].

Dado un conjuntos de datos en R^n, el objetivo del SOM es encontrar un conjunto de vectores $m_i \in R^n$, a los que se llama modelos o prototipos, asociados uno a uno con los nodos de la malla y distribuidos de tal forma que modelos cercanos en R^n son cercanos en la malla, estos vectores modelo se pueden pensar como promedios ponderados locales de los objetos en el espacio de datos.

El algoritmo para calcular el mapeo del SOM sigue pasos bastantes simples, donde el paso básico consiste en la actualización de los modelos ante un nuevo dato. La actualización está dada por la ecuación

$$m_k(s+1) = m_k(s) + \alpha(s)\eta(s,k,v)(x(t) - m_k(s)) \tag{2.13}$$

donde v es el vector modelo asociado a la unidad más similar a x(t) al tiempo s, $\eta(s,k,v)$ es la función de vecindad, y $\alpha(s)$ el coeficiente de aprendizaje. El algoritmo se puede ver en 1.

Algorithm 1 Algoritmo SOM

Sea $x_1, \ldots, x_m \in R^n$, una secuencia de datos
Iniciamos aleatoriamente los vectores prototipo $m_1, ..., m_r$
Iniciamos el radio efectivo de la vecindad σ y el coeficiente de aprendizaje α, tal que $\alpha < 1$
for $s \leftarrow 1, s_{max}$ **do**
  for $t \leftarrow 1, m$ **do**
    for $k \leftarrow 1, r$ **do**
      $m_v = argmin_j \|x(t) - m_j(s)\|$
      $m_k(s+1) = m_k(s) + \alpha(s)\eta(s,k,v)(x(t) - m_k(s))$
      Disminuir α y el radio efectivo de la función de vecindad η
    end for
  end for
end for

Originalmente, la función de vecindad se planteó como una función escalonada tal que $\eta(s,k,v) = 1$ si u_k está en la vecindad de u_v, es decir si $|u_k - u_v| \leq \sigma(s)$ y 0 en el resto de la malla, sin embargo, en la práctica se utiliza una función de tipo gausiano como $\eta(s,k,v) = e^{\frac{\|w_k(s) - w_v(s)\|}{2\sigma(s)^2}}$.

El SOM, genera un mapeo que preserva topología de los datos de entrada, y puede ser visto como una gráfica de similitud. Un problema del SOM es que el mapeo es a una malla rígida, por lo que, en general, la proximidad entre los modelos no puede ser bien representada sólo con la distancia entre los nodos de la malla, ya que, por ejemplo, dos nodos vecinos podrían en realidad ser distantes en el espacio de los datos. Para dar una imagen de cuan lejanos o cercanos son dos nodos vecinos, una técnica son las matrices de distancias de las cuales la más conocida es la matriz de distancia unificada (matriz-U), con la cual, utilizando gradientes de grises, se representa de forma cualitativa la distancia entre los modelos en el espacio de datos original. Cuando esto no es sufi-

ciente, y se requiere una representación de los datos que aproxime mejor la estructura de distancias de los datos, es decir, que la represente cuantitativamente, existe una extensión del SOM conocida como ViSOM. El algoritmo del ViSOM sí busca mantener las distancias, para así conseguir un escalamiento métrico [62].

En general, el SOM tiene la cualidad de ser un método poderoso y flexible, tolerante al ruido, además de simple, y fácil de explicar y visualizar [57]. Aunque, en contraste con la sencillez del algoritmo y la presentación de los resultados, la teoría matemática resulta bastante complicada y únicamente el caso de una dimensión se ha analizado completamente [35]. Esto es una desventaja que tiene en comparación con otros métodos y que genera críticas al modelo.

Aunque los métodos de las secciones anteriores también son considerados aprendizaje no supervisado[30], el SOM pertenece a una clase distinta. En particular, porque no tiene una función de ajuste a la que siga [64], esto por un lado tiene la desventaja de que complica la construcción de una teoría matemática que explique totalmente el funcionamiento del SOM, pero, por otro lado, implica una flexibilidad inherente que otros métodos no tienen, además de una simplicidad que resulta bastante atractiva.

Algo en lo que se debe ser cuidadoso con el SOM es en la definición de los parámetros del algoritmo, ya que esto puede afectar sensiblemente su desempeño, por lo que a veces se requiere de un proceso de ajuste de parámetros- como el tamaño de la malla- para definirlos de manera *ad hoc* [18].

Finalmente, una ventaja que el SOM tiene sobre los métodos anteriores es que es un algoritmo adaptativo, que genera un mapeo explícito, que permite aplicarlo a nuevos datos sin necesidad de recalcularlo.

SOM por lotes

El SOM por lotes es una modalidad donde los vectores modelo se actualizan simultáneamente, utilizando todos los datos de la muestra al mismo tiempo. La ventaja de esta versión es que es bastante más rápido y sus resultados similares al original.

En este caso, el paso de actualización lo que hace es asignar directamente a cada modelo una especie de promedio de los datos asignados a su vecindad, como se puede observar en 2.14

$$m_k(s+1) = \frac{\sum_j \eta(s,k,v_{x_j}) x_j}{\sum_j \eta(s,k,v_{x_j})} \tag{2.14}$$

donde v_{x_j} es el vector modelo más cercano a x_j, y la función de vecindad, al igual que antes, reduce su radio con el tiempo.

2.3.5. Métodos para datos no vectoriales

2.3.6. SOM para datos no vectoriales

El SOM para datos no vectoriales es una adaptación de la versión anterior. Como en este caso no existe una aritmética en el espacio de los datos, ni una distancia, el algoritmo tiene que plantearse diferente. En particular, como se verá, los modelos siempre son elementos del conjunto de datos.

Para definir un SOM con datos no vectoriales, es necesario hacer algo equivalente a calcular el promedio de un conjunto de datos. Para esto se define la media generalizada.

Definición 2.3.1 *Dado un conjunto S de objetos y una medida de distancia d entre ellos, se define la media generalizada de S como*

$$M(S) = arg\ min_{x \in S} \sum_{y \in S} d(x,y) \tag{2.15}$$

En caso de que se tenga una medida de similitud, es decir, para la cual valores altos representen mucha similitud, en lugar del mínimo habría que tomarse el máximo. La media generalizada es central para el SOM no vectorial, por lo que se le suele llamar Median SOM.

Como se mencionó antes, una característica importante de este tipo de SOM es que los modelos son elementos de la muestra, esto se define explícitamente en la definición de la media generalizada pidiendo que la media sea un elemento del conjunto. Una de las consecuencias de esta característica es que se reduce la complejidad del algoritmo, ya que el SOM sólo requiere conocer las similitudes entre los objetos y no necesita calcular ningún otra a lo largo del entrenamiento.

El algoritmo funciona como sigue, para cada modelo m_i en la malla del SOM, definimos N_i como el conjunto de modelos en la vecindad de m_i, si denotamos por N_i^{-1} al conjunto de objetos x en S que son mapeados en N_i, es decir, tales que

$$arg\ min_{m \in S} d(x, m) \in N^{-1} \tag{2.16}$$

entonces el paso de actualización de los modelos está dado por

$$m_i = M(N_i^{-1}) \tag{2.17}$$

este paso, como en el caso del SOM por lotes, se realiza simultáneamente para todos los modelos, a su vez también el radio que define N_i se puede ir reduciendo.

Dado que los modelos del SOM se obtienen de un conjunto finito de posibilidades es probable que existan dos modelos idénticos, por lo que hay que definir una regla de desempate cuando se calcula el modelo más afín a un objeto dado. La solución es considerar los demás modelos alrededor de cada uno de los modelos empatados y dar la victoria a aquel para el cual la suma de las distancias de su vecindad al objeto es menor.

Capítulo 3

Análisis de datos espaciales

Muchos fenómenos en la naturaleza y en nuestra sociedad tienen un componente espacial. Algunos ejemplos de fenómenos que se pueden analizar desde esta perspectiva son la agricultura, los ecosistemas, las epidemias, los flujos migratorios, o el crecimiento de bacterias en una solución. Además, el aumento en el uso de dispositivos con capacidad para conocer su ubicación, permite que día a día se generen más datos espaciales, datos que van desde el ámbito científico –datos de colecciones de especies, datos climáticos por satélite– hasta el casual –por ejemplo, mediante aplicaciones en teléfonos móviles como Foursquare, una aplicación móvil donde los usuarios registran dónde comen y comentan la calidad del sitio. En conjunción con esto, también se están creando grandes bases de datos abiertas al público. Un ejemplo es la base de datos espaciales GBIF (Global Biodiversity Information Facility, www.gbif.org), otros son los portales de datos públicos gubernamentales como www.data.gov, www.data.gov.uk (portales de datos de los gobiernos de Estados Unidos y del Reino Unido, respectivamente). La consecuencia de esto es que contamos con una inmensa cantidad de datos abiertos a nivel mundial que abarcan un espectro de temas enorme. Todos estos datos contienen información implícita sobre las interacciones que dan forma a nuestro mundo y sociedad. Este conocimiento es de interés tanto para expertos de área como para el consumidor

de información casual –ciudadanos interesados en temas urbanos, entusiastas de la ecología, campañas de salud–. El reto es desarrollar mejores herramientas para transformar esa información en conocimiento, en formas que sean útiles tanto para expertos como para usuarios no especializados en el análisis de datos. Para eso requerimos codificaciones visuales intuitivas, que logren condensar grandes cantidades de datos y presentar la información de forma que sea más fácil de procesar, para así lograr una especie de democratización de la información y explotar mejor el potencial de este conocimiento.

Retos como este han generado nuevos esfuerzos en la comunidad científica, con el fin de estudiar, diseñar, evaluar métodos para procesar esta información, y presentarla en formas que se ajusten al tipo de tareas de análisis visual para las que el cerebro humano está mejor capacitado. Como efecto de este esfuerzo surgió un nuevo término, Analítica Geovisual [53], que se define como "la ciencia del razonamiento analítico y toma de decisiones con información geoespacial, facilitado por interfaces visuales interactivas, métodos computacionales, representaciones de construcción del conocimiento y estrategias de gestión"[15]. Dos de los campos que involucra la Analítica Geovisual son minería de datos espaciales y visualización. Para avanzar en el programa y objetivos de ésta, necesitamos integrar técnicas de minería de datos y análisis estadístico con herramientas que apoyen el pensamiento visual. Esta sección sienta las bases para una metodología cuyo objetivo va en esa dirección. En este caso, dirigida al análisis visual de datos geográficos representados por grandes colecciones de variables espaciales booleanas, utilizando estadísticas de co-ubicación. Es importante subrayar que, aunque nos centremos en datos geoespaciales, la metodología que presentamos no se restringe a datos geo-referenciados, ya que lo único que asume es que se tienen variables distribuidas en un espacio de coordenadas continuas. Visto desde la nomenclatura de Bertin, hablamos de datos cuyos componentes de referencia son continuos, y que en el contexto de Analítica Geovisual se restringen usualmente a dos (latitud, longitud) o tres (latitud, longitud y tiempo; o latitud, longitud, altitud) dimensiones, y en raras ocasiones a cuatro dimensiones.

3.1. Diferencias con la minería de datos tradicional

La minería de datos espaciales es en cierto sentido más complicada que la minería de datos tradicional. Generalmente, cuando se trata de minería de datos tradicional, las relaciones de co-ocurrencia son explícitas. Por ejemplo, para analizar patrones de compra en una tienda, las asociaciones entre productos se pueden inferir de los tickets de compra, donde una co-ocurrencia entre dos productos se da si están en el mismo ticket. Una de las complicaciones en minería de datos espaciales es que las relaciones de co-ocurrencia entre variables usualmente no están representadas explícitamente. Relaciones como cercanía, contención o superposición, requieren que uno calcule la distancia entre dos puntos o determinar si un polígono está contenido en otro, o determinar si dos polígonos se intersecan. Para resolver este problema existen dos filosofías, utilizar algún método para materializar la relación de forma explícita en los datos, o utilizar métodos estadísticos que puedan lidiar con este tipo de relaciones. En nuestro caso utilizaremos la primera.

Otra característica común en datos espaciales es la autocorrelación espacial, o lo que en algunos textos de ciencia de la información geográfica se llama la "Primer ley de geografía de Tobbler", que enuncia: "Todo está relacionado con todo, pero cosas cercanas están más relacionadas entre sí que cosas lejanas". Esta propiedad tiene consecuencias que hacen menos apropiados métodos de estadística clásica. Por ejemplo, no siempre es adecuado considerar que un fenómeno registrado como una distribución de puntos en el mapa se puede tratar directamente como un proceso aleatorio e independiente, que se distribuye con probabilidad uniforme en el espacio de estudio. Debido a la auto-correlación el fenómeno tenderá a formar cúmulos de puntos lo cual no puede ser modelado como un proceso de Poisson. Aunque el problema de la autocorrelación está muy bien ubicado, no existe una solución estándar para analizar datos con esta característica, y en general se dice que aunque no se aplique un método para lidiar con la autocorrelación, es importante que el analista este consiente de su existencia.

3.2. Análisis visual de datos espaciales

El análisis de datos multivariados es complicado, y si añadimos el componente espacial se complican más las cosas. Principalmente porque se aumentan una o dos dimensiones al espacio de referencia de los datos, es decir, mientras que en bases de datos tradicionales las relaciones están representadas en los registros mismos, en bases de datos espaciales las relaciones se extraen del espacio de referencia. Debido a su complejidad, usualmente es necesaria más de una perspectiva visual. Por ejemplo, en [27] los autores presentan una metodología para el análisis visual de interacciones espaciales entre localidades de Estados Unidos, donde utilizan tres tipos de visualización[*]: la de relaciones multivariadas (diagramas de coordenadas paralelas); la geográfica (mapas de flujos); y la de redes (con una red que representa las interacciones espaciales como aristas y las localidades como nodos). Cada una de estas visualizaciones es parte del sistema de representación porque es mejor para mostrar cierto tipo de patrones en los datos. La vista geográfica permite estudiar las correlación entre las variables y el espacio geográfico, y si las relaciones entre unas y otras se ven afectadas por la geografía y por cercanía; la de relaciones multivariadas nos permite detectar si hay patrones que implican relaciones entre varias variables; y la red de interacciones permite simplificar el espacio para ver que localidades se asocian con cuales. El problema es que cuando el número de variables es muy grande, entonces, es mucho más complicada la búsqueda de relaciones entre variables y muchos de estos métodos no se escalan.

En ciencia de la información geográfica, probablemente el mejor ejemplo de la utilidad de las redes como representación de relaciones entre objetos, son las redes espaciales. Las redes espaciales han jugado un papel importante desde finales del siglo XIX hasta nuestros días. Se les llama redes espaciales a redes de infraestructura, de comunicaciones, y de transporte [13], o a redes que

[*]En información geográfica se conoce como interacciones espaciales cuando dos localidades geográficas tienen algún tipo de intercambio, por ejemplo, si población de una localidad migra a otra entonces hay una interacción entre estas localidades

mapean datos de interacciones espaciales [47]. Uno puede encontrar en la literatura ejemplos de redes espaciales utilizadas para visualización [27] o para definir medidas y analizar sus propiedades topológicas y geométricas [60]. Por ejemplo, una red de carreteras se puede representar mediante nodos que representan cruces de carreteras y aristas que representan las carreteras. Las propiedades topológicas de la red pueden ser usadas como ayuda en la gestión de crisis, usando medidas de teoría de gráficas como el grado de intermediación de los nodos para detectar puntos altamente vulnerables [1]. Otro ejemplo se encuentra en [22] donde los autores definen un modelo generador de redes derivado de redes espaciales observadas, y este modelo es utilizado para explicar el origen de algunas propiedades comunes en redes de origen geográfico.

3.3. Análisis de correlación espacial entre variables espaciales booleanas

Dado un conjunto de variables espaciales, es importante conocer cómo se relacionan entre sí a partir de sus distribuciones espaciales. Si contamos con N conjuntos de puntos, donde cada uno de ellos corresponde a una variable específica, la pregunta es ¿existe alguna asociación entre algunas de estas variables? y de ser así ¿cómo podemos medir esto?

El primer paso es definir cuándo dos eventos se relacionan. Supongamos que en un estudio se obtuvieron datos de ingreso anual por casa y que se eligieron un número de casas al azar en las cuales se preguntó el ingreso anual. Supongamos también que tenemos datos de otro estudio donde se obtuvo el número de habitantes por hogar, donde también se eligió un cierto número de casas al azar. Cada dato está asignado a la dirección de la casa. Como no se cubrieron todas las casas de México en ninguno de los estudios, probablemente habrá pocas para las que se tengan ambos datos, entonces ¿cómo podemos medir si el número de habitantes en una casa nos da alguna información sobre el

ingreso anual de ese hogar? (y viceversa). Una solución es generar unidades de área que contengan más de una casa y calcular las estadísticas sobre estas nuevas unidades, es decir, agregamos varias casas por zonas y consideramos los datos que se recolectaron en una zona como datos que ocurrieron en el mismo sitio. Por ejemplo, podemos considerar casas por colonia, promediar los valores de las variables, y hacer el análisis de correlación al nivel de colonias. Por supuesto, podríamos elegir otra escala más fina (por manzana), o más gruesa (por municipio o delegación). Esta decisión claramente afecta nuestros resultados, y es un tema que sigue abierto a propuestas sobre como lidiar con él de manera estándar. Cabe resaltar, que el ejemplo anterior usa un tipo de división del espacio irregular, otra opción, como veremos más adelante, es usar una división regular, como una rejilla.

Un tipo de variable que es común tanto en datos espaciales, como en los no espaciales, son las variables booleanas. Por ejemplo, cualquier variable que informa sobre la presencia de alguna característica. En otros casos, estas pueden ser construidas a partir de variables que toman valores en un conjunto más grande, por ejemplo, a partir de datos de temperatura se puede construir una nueva variable que indique en cada punto si la temperatura es mayor a 20^o. Este tipo de variables serán el objeto de nuestro estudio.

3.3.1. Modelos de co-ubicación local

En la discusión anterior se mencionó que se podría analizar el conjunto de datos espaciales utilizando divisiones políticas del territorio para agregar los datos, y que otra opción es utilizando una rejilla regular donde dos eventos coinciden en el espacio si caen en la misma celda. Aunque en ciertos casos utilizar la división por municipios sería la decisión natural, es claro que en otros esto no sería adecuado y esto dependerá no sólo del tipo de datos, sino de la pregunta que se busca responder. Al tipo de estrategia que se utiliza para decidir como agregamos los datos le llamaremos modelo de co-ubicación.

Un tipo de análisis que se hace a partir de modelos de co-ubicación es la detección de patrones de co-ubicación. Los patrones de co-ubicación sirven para inferir reglas que hablan sobre la presencia de un tipo de objeto o característica a partir de la presencia de otros tipos de objeto o característica. Por ejemplo, un patrón de co-ubicación podría ser que registros de agua contaminada aparecen cerca de registros de población con un alto índice de enfermedades estomacales. Los patrones de co-ubicación son conjuntos de características espaciales booleanas que aparecen cerca frecuentemente y por lo general son utilizados para derivar reglas de co-ubicación, las cuales son reglas de inferencia, es decir, reglas del tipo "si hay instancias de las características f_1 y f_2 en una localidad entonces es probable que haya instancias de las características f_3 y f_4 en la vecindad". Entonces, dado un conjunto de características espaciales booleanas C, un patrón de co-ubicación es un subconjunto de características en C que aparecen cercanas frecuentemente, mientras que una regla de co-ubicación es una implicación lógica. En el ejemplo del párrafo anterior podríamos obtener la regla 'si hay agua contaminada entonces es probable que exista población con un índice alto de enfermedades estomacales en la zona', o en el otro sentido, 'si una población tiene un alto índice de enfermedades estomacales entonces es probable que el agua de la zona esté contaminada'.

Como mencionamos en el capítulo anterior, la forma en la que uno estructura los datos, no depende sólo de los datos, depende de las preguntas que buscamos responder. En este sentido existen distintos esquemas para construir co-ubicaciones. Los esquemas más reconocidos son: basado en ventanas, basado en eventos, y característica de referencia. De acuerdo con [48] cada modelo está orientado a aplicaciones distintas. El modelo de característica de referencia es importante para estudios enfocados en una característica espacial booleana f que buscan encontrar co-ubicaciones con otras características espaciales booleanas que ayuden a predecir la distribución de f. El basado en ventanas es utilizado en estudios enfocados en parcelas de tierra. En este caso, el objetivo sería obtener la probabilidad de que un conjunto de n características estén presentes en un pedazo de tierra dado que esa área de tierra cumple con ciertas m características. Por último, el modelo basado

en eventos es relevante en casos donde se tienen muchos tipos de características espaciales y estamos interesado en encontrar subconjuntos de características que tienden a estar presentes en algún tipo de evento de interés. Para cada uno de estos esquemas se tienden a definir tipos específicos de patrones de co-ubicación, por ejemplo en el esquema de característica de referencia, los patrones de co-ubicación siempre contendrán a la característica de referencia. Al tipo de patrón de co-ubicación le llamaremos modelo de co-ubicación local.

A su vez, dado un modelo de co-ubicación local, se puede definir de distintas maneras cuándo se realiza un patrón de co-ubicación, utilizando una rejilla uniforme, una teselación, o definiendo un radio de influencia. En este trabajo se define el modelo de co-ubicación utilizando rejillas uniformes, de tal modo que dos características espaciales booleanas tienen una co-ocurrencia por cada celda que contenga instancias de ambas.

3.3.2. Épsilon para datos espaciales

Una vez definido el modelo de co-ubicación podemos construir una estadística de diagnostico sobre las dependencias entre variables. En este caso utilizaremos ε 2.2.3. Si tenemos dos características espaciales X y Y, calculamos el número de co-ocurrencias en dos pasos: Primero, definimos la resolución de la rejilla (ej. celdas de 1 Km por lado); después aplicamos la rejilla y contamos en cuántas celdas se encuentran tanto X como Y.

Para calcular ε, necesitamos N_t, el total de celdas en nuestro universo, N_X el número de celdas ocupadas por X, N_Y el número de celdas ocupadas por Y y N_XY, el número de celdas ocupadas por ambas. Dichos datos nos permiten calcular

$$\varepsilon_{X,Y} = \frac{N_Y(\frac{N_{X,Y}}{N_Y} - \frac{N_X}{N_t})}{\sqrt{N_Y \frac{N_X}{N_t}(1 - \frac{N_X}{N_t})}} \tag{3.1}$$

Como veremos a continuación, los resultados de ε –y de cualquier estadística derivada de un mo-

delo de co-ubicación– dependen de cómo se agreguen los puntos, en este caso la resolución, la posición y la rotación de la rejilla son parámetros que afectan los resultados.

3.3.3. Agregación de datos y el problema de la unidad de área modificable

La agregación de datos puede ocurrir en dos momentos: al momento de captura de los datos y en la etapa de preparación de los datos para el análisis. Un ejemplo de agregación de datos durante la captura de estos es en imágenes satelitales, donde un pixel corresponde en realidad al promedio en un área. En el procesamiento de datos previo al análisis la agregación ocurre en situaciones en las que se debe elegir un modelo de co-ocurrencia para inferir las relaciones entre tipos de eventos. Esta agregación de datos produce un sesgo en el análisis que es conocido como el problema de la unidad de área modificable (*Modifiable areal unit problem*, MAUP).

El primer paso en nuestra metodología es definir el modelo de co-ocurrencia con una rejilla. Por lo tanto, nuestro primer problema, y el de cualquier metodología que usa alguna noción de discretización espacial, es el problema de la unidad de área modificable (MAUP). Uno de los primeros reportes relacionados con el MAUP se encuentra en [23], una publicación de hace mucho tiempo. Sin embargo, el problema fue ignorado por décadas, y hasta finales de los 70 se empezó a considerar nuevamente, probablemente debido a que los avances en computación hicieron posible la implementación de nuevos algoritmos para generar posibles soluciones [45] y dejó de verse como un problema irresoluble.

El MAUP sigue siendo un tema de investigación activo y sigue como pregunta abierta si se pueden desarrollar métodos automáticos para detectar escalas apropiadas de muestreo [1]. En nuestra metodología, el MAUP es consecuencia de la resolución de la rejilla, la cual es una versión relativamente más simple que el caso general donde la división no es regular y se pueden elegir los trazos de la zonificación. Un ejemplo de agregamiento de datos con áreas irregulares son los AGEB (Área

GeoEstadística Básica) que utiliza el INEGI [*].

El MAUP tiene dos componentes: de escala y de zonificación. Por un lado la escala de la división, por ejemplo pasar de colonias a estados, o aumentar el tamaño de las celdas en la rejilla. Por otro, se puede mantener la escala y cambiar la forma, la orientación o la posición de las zonas, por ejemplo el INEGI puede redefinir los AGEB, o si la división se define con una rejilla esta se puede trasladar o rotar y los datos se agregarán diferente.

En los párrafos siguientes analizaremos los efectos de la resolución y argumentaremos que hay resoluciones que son mejores, lo cual depende de los puntos en el conjunto de datos. También mostraremos que la elección de la resolución puede tener un fuerte impacto en el análisis.

El efecto de la elección del tamaño de las celdas en una rejilla se puede apreciar si consideramos los casos extremos. En una rejilla compuesta por una sola celda que cubre completamente nuestra área de análisis, todo cae en la misma canasta y por lo tanto todo esta igualmente relacionado con todo, pues todas las variables tendrán una y sólo una co-ocurrencia, por lo que no obtenemos ninguna información. En el otro caso, tenemos una resolución tan fina que nada coincide con nada y tampoco obtenemos información. Podemos pensar en dos estrategias para elegir una rejilla de forma automática: una se deriva del efecto que tiene la resolución en el tamaño efectivo de la muestra y la otra del efecto directo en la medida de significatividad estadística de nuestra elección. En la primera, la lógica es que como estamos haciendo un análisis estadístico y prueba de hipótesis es natural intentar aprovechar al máximo las muestras disponibles, es decir, maximizar el número de celdas con coincidencias.

El primer paso en el proceso es mapear los puntos en las celdas de la rejilla, ya que el análisis estadístico se hace al nivel de celdas y no de eventos. Si una variable aparece en una celda se cuenta una vez, sin importar si aparece una o cien veces, por lo tanto si la rejilla es tal que en promedio muchos eventos ocurren por celda, el tamaño efectivo de nuestras muestras puede reducirse con-

[*] http://mapserver.inegi.gob.mx/geografia/espanol/prodyserv/cartocen/cartocen.cfm?c=334

siderablemente. Por ejemplo, para una distribución aleatoria de eventos, si en promedio el número de eventos por celda es cuatro, entonces una reducción del tamaño de las celdas por un factor de 2 probablemente producirá celdas donde el número de eventos esperado por celda es más cercano a uno. En otras palabras, el número de celdas debería escalarse naturalmente en proporción al número de eventos. En el caso de co-ocurrencias para un conjunto finito de eventos, si recurrimos a celdas cada vez más pequeñas es claro que el número de co-ocurrencias tenderá a cero. Por otro lado, si recurrimos a celdas muy grandes terminaremos con una co-ocurrencia entre cualquier par de características, ya que todos los eventos estarán en una celda.

Si consideremos dos variables espaciales booleanas X y Y, y el número de co-ocurrencias entre ellas como una función N_c de la resolución de la rejilla r, entonces nuestra función es tal que

$$max(N_c(r)) \leq min(\#X, \#Y) \tag{3.2}$$

Por otro lado, si la rejilla la hacemos muy fina

$$Lim_{r \to 0} N_c(r) = 0 \tag{3.3}$$

Finalmente, si una celda contiene todos los puntos, entonces

$N_c(r) = 1, r \geq L_M$, donde L_M es la distancia máxima entre cualesquiera dos puntos en nuestra muestra (3.4)

La ecuación 3.2 expresa que a lo más el número de co-ocurrencias es el número de eventos en el conjunto más pequeño de los dos; la ecuación 3.3 expresa que si la resolución es muy fina no tendremos co-ocurrencias; y por último, la ecuación 3.4 que si una celda cubre completamente nuestros conjuntos de puntos, entonces tenemos una y sólo una co-ocurrencia. Si para elegir la re-

solución de la rejilla seguimos una estrategia basada en maximizar el tamaño efectivo de la muestra, entonces nuestro objetivo es maximizar el número de eventos de interés, en este caso el número de co-ocurrencias entre X y Y.

3.4. Selección de rejilla para datos espaciales

3.4.1. Exploración de los efectos de la resolución de rejilla

En esta sección exploraremos los efectos que tiene la resolución de la rejilla en el resultado del análisis de correlación entre variables espaciales. Parte de lo que se presenta a continuación está incluido en la publicación [51]. Comenzamos el análisis con datos generados artificialmente, de tal forma que sabemos como se generaron las distribuciones y sabemos cuando hay dependencia y cuando no. En el segundo análisis, utilizamos datos de colectas de mamíferos y especies de lutzomyia. En ambos casos el análisis consiste en ver como varía el diagnóstico estadístico con respecto a los cambios de resolución. El objetivo es entender las características que distinguen una rejilla adecuada para detectar si hay dependencia entre variables, desarrollar intuición para interpretar los resultados de los diagnósticos estadísticos, y recalcar que la rejilla que se utiliza para el análisis por sí sola da información sobre el tipo de asociaciones que se están detectando.

Dado que en estadística, entre más grandes sean las muestras de datos, más certeza tenemos en los resultados, y en el diagnóstico de correlación entre distribuciones espaciales parte de la muestra son las coincidencias que hay entre las variables, es de interés saber como se comporta el número de coincidencias. Y, como explicaremos más adelante, una estrategia para aumentar la confianza en nuestro análisis, es tratar de maximizar el número de coincidencias. Dicho de otro modo, maximizar el número de celdas ocupadas por ambas variables.

La resolución de una rejilla es como el foco de un lente, el cual necesitamos ajustar para ver con

claridad a la escala en que se dan las relaciones entre dos fenómenos. En ese sentido, necesitamos calibrar la resolución de la rejilla para que 'enfoque' a la escala adecuada. Esto nos lleva a plantearnos dos preguntas, la más básica es ¿se puede? (aunque sea aproximadamente), y si se puede, la pregunta obvia es ¿cómo?. En esta sección argumentamos que sí se puede hacer algo y proponemos una forma de hacerlo, aunque como es común el cómo tiene muchas respuestas, en nuestro caso presentamos lineamientos para hacer la elección de rejilla de manera automática. Para una discusión más extensa sobre el problema de elección de rejilla, y otros métodos de agrupamiento de datos espaciales, el artículo "The modifiable areal unit problem" [45] es el texto que resucitó la discusión de este problema, el cual llevaba décadas siendo ignorado. Otra pregunta que nos podemos hacer sobre la escala es ¿qué es lo que detecta?, no buscamos responder esta pregunta, pero vale la pena tenerla en mente. En el caso de especies uno podría interpretar esta relación más como una relación evolutiva, es decir que detecta que dos organismos evolucionaron para volverse parte de un mismo ecosistema.

Consideramos primero el caso más simple: dos variables espaciales booleanas, aleatorias, independientes, que se distribuyen uniformemente en un área. Para ello, generamos dos conjuntos de puntos distribuidos aleatoriamente en un cuadrado, el primer conjunto de puntos representa una variable X y el segundo una variable Y. Generamos un tercer conjunto de puntos X' cuya distribución depende de Y. Con esto tenemos dos parejas X, Y y X', Y, y podemos comparar las diferencias entre el caso de variables independientes y el de variables dependientes. El siguiente paso es contar el número de coincidencias, y calcular el valor de ε para cada rejilla, de una sucesión de rejillas que subdividen el área a resoluciones cada vez más finas. La sucesión de refinamientos de la rejilla se produjo empezando con una sola celda que abarca el área completa y aumentando la siguiente resolución para obtener una celda más por lado, es decir, tenemos una sucesión de rejillas regulares que abarcan la misma área, donde la rejilla r_k, tiene $k \times k$ celdas y la rejilla r_{k+1} tiene $k + 1 \times k + 1$. Además de considerar los casos de variables independientes y dependientes, analizamos como afecta la razón

de puntos entre una distribución y la otra. Para esto realizamos el análisis para el caso en que los conjuntos tienen el mismo número de puntos, y luego repetimos el proceso con conjuntos X y X' más grandes que el conjunto Y. Los resultados se pueden apreciar en 3.1.

En 3.1 podemos ver que al inicio, conforme la resolución se hace más fina, la tendencia es que aumente el número de coincidencias, pero después de un punto se revierte y el número de co-ocurrencias empieza a decrecer aproximándose a cero. Por otro lado, en ambos casos parece haber un máximo global, y la gráfica se aproxima a una función cóncava, aunque con algo de ruido. Esto nos sugiere que si el objetivo es maximizar el número de co-ocurrencias, y así maximizar el tamaño efectivo de la muestra de eventos de interés, se podría lograr con métodos de optimización simples, como descenso de gradiente.

Dadas dos distribuciones con n puntos cada una, el máximo de coincidencias que puede haber es n, para que se dé este máximo no puede haber más de un punto por celda, y en todas las celdas ocupadas deben haber puntos de las dos distribuciones, esto es muy difícil que ocurra, y en el caso de variables independientes aún usando la rejilla que maximiza el valor esperado de celdas con coincidencias éste está lejos de ser igual al número de puntos. Esta reflexión, aunque simple, nos da algunas ideas sobre la función que queremos optimizar.

Dentro de las cosas que necesita cumplir la rejilla es que el valor esperado del número de puntos en una celda no sea mayor que uno, en realidad, debiera ser menor que uno, ya que si es uno, esto quiere decir que hay un 50 % de probabilidad de que haya más de un punto en una celda arbitraria. Si nuestras distribuciones de puntos son del mismo tamaño este principio parece bastar, sin embargo, cuando las muestras de las variables espaciales son dispares en el número de instancias y una es considerablemente más grande que la otra, resulta evidente otro tipo de problema, que con usando una rejilla resulte que una variable este presente en todas las celdas. Es importante que la variable con más puntos no cubra demasiadas celdas, ya que si el máximo de coincidencias se alcanza cuando la variable cubre casi el total de las celdas, entonces no se detectaría relación

alguna, ya que tendríamos $P(C) \approx P(C|X)$ o en el caso extremo $P(C) = P(C|X)$. Por lo tanto, debemos pedir que el valor esperado del número puntos de ambas variables sea estrictamente menor que uno. Visto de esa manera, en general, lo que buscamos es que la muestra contenga tanto casos donde se encuentra la variable de referencia como casos donde no.

Si observamos las diferencias entre la gráfica que corresponde a variables independientes (azul) y la que corresponde a variables correlacionadas (rojo), veremos que para la segunda se requiere una resolución más fina para acercarnos al máximo, esto es simplemente consecuencia de que una variable se distribuye más densamente alrededor de la otra. A mayor dependencia de X con Y se requiere mayor refinamiento de la rejilla para obtener el máximo.

	N_x	N_y	N_{xy}	N	ε
1	1	1	1	1.00	0.00
2	4	4	4	4.00	0.00
3	16	16	16	16.00	0.00
4	64	49	49	64.00	0.00
5	227	82	75	256.00	0.80
6	408	96	37	1024.00	-0.26
7	473	98	9	4096.00	-0.73
8	495	99	2	16384.00	-0.58
9	499	99	0	65536.00	-0.87
10	499	100	0	262144.00	-0.44
11	500	100	0	1048576.00	-0.22
12	500	100	0	4194304.00	-0.11
13	500	100	0	16777216.00	-0.05

Tabla 3.1: X, Y independientes. Tabla que muestra los resultados de corridas con resoluciones cada vez más finas, con 500 puntos de X y 100 de Y, ambas con distribución aleatoria uniforme

En resumen, la rejilla óptima va a depender del tamaño de las muestras, del grado de dependencia estadística entre las variables, y del radio de influencia que tienen las ocurrencias de una sobre la otra. Todos estos son parámetros que determinan la resolución a la que el número de co-ocurrencias se maximiza y el grado de sensibilidad del diagnostico estadístico ante cambios de resolución.

	N_x	N_y	N_{xy}	N	ε
1	1	1	1	1.00	0.00
2	4	4	4	4.00	0.00
3	16	16	16	16.00	0.00
4	49	49	48	64.00	3.54
5	86	82	81	256.00	12.50
6	104	96	94	1024.00	28.47
7	116	98	96	4096.00	56.77
8	154	99	94	16384.00	96.94
9	211	99	88	65536.00	155.56
10	305	100	78	262144.00	228.46
11	418	100	42	1048576.00	210.20
12	473	100	18	4194304.00	169.40
13	494	100	2	16777216.00	36.80

Tabla 3.2: X dependiente de Y. Tabla que muestra los resultados de corridas con resoluciones cada vez más finas, con 500 puntos de X y 100 de Y con distribución uniforme. Donde la distribución de X depende de la de Y. [EXPLICAR COMO]

En realidad, para el caso de variables espaciales independientes podemos hacer algo más que simulaciones, se puede expresar analíticamente el número esperado de coincidencias. El problema creo que es más simple pensarlo con canicas y cubetas. Si tenemos un costal con k_r canicas rojas y k_a canicas azules y los vaciamos en n cubetas, ¿cuál es el número esperado de cubetas con al menos una canica azul y una roja?. La probabilidad de que una canica caiga en una cubeta i es $\frac{1}{n}$, por lo tanto la probabilidad de que la cubeta no contenga la canica es $\frac{n-1}{n}$, de lo que se sigue que la probabilidad de que no contenga ni una de las k_a canicas azules es $(\frac{n-1}{n})^k_a$. Sea a_i la variable aleatoria tal que $a_i = 1$ si la cubeta i contiene al menos una canica azul y $a_i = 0$ si no es así. Entonces

$$P(a_i) = 1 - \left(\frac{n-1}{n}\right)^{k_a} \tag{3.5}$$

análogamente para la variable r_i, que indica si la cubeta i contiene al menos una canica roja, $P(r_i) = \left(1 - \left(\frac{n-1}{n}\right)^{k_r}\right)$. Si definimos c_i la variable aleatoria tal que $c_i = 1$ si la cubeta i contiene

canicas de ambos colores, y $c_i = 0$ en caso contrario. Entonces, dado que dónde cae una canica es independiente de dónde cayeron las demás, tenemos que

$$P(c_i) = P(a_i)P(r_i) = \left(1 - \left(\frac{n-1}{n}\right)^{k_a}\right)\left(1 - \left(\frac{n-1}{n}\right)^{k_r}\right) \tag{3.6}$$

Sea C el número de celdas con coincidencias, $C = \sum_{i=1}^{n} c_i$, entonces el valor esperado de C es

$$E(C) = E(\sum_{i=1}^{n} c_i) = \sum_{i=1}^{n} E(c_i) = n\left(1 - \left(\frac{n-1}{n}\right)^{k_a}\right)\left(1 - \left(\frac{n-1}{n}\right)^{k_r}\right) \tag{3.7}$$

nótese que las c_i no son independientes entre sí, ya que si sabemos que un número de canicas cayeron en una cubeta dada, esto cambia la probabilidad de que las otras cubetas contengan una canica. Sin embargo, podemos abrir la suma gracias a la linealidad del valor esperado aún cuando las variables son dependientes. Además, sabemos que $E(c_i) = P(c_i)$.

En 3.2 se presentan gráficas de $E(C)$ para una distribución X de 50 puntos y distribuciones Y de 100, 50, 25, 10 y 5 puntos.

Si una variable tiene más instancias que la otra, la menor determina el máximo de co-ocurrencias que se pueden obtener. En este caso, un requisito es que la probabilidad de que haya más de un punto por celda de la distribución más chica sea cercana a cero, y no importa si la otra variable cumple esta condición. Pero, como ya mencionamos, si buscamos maximizar la medida de asociación estadística, entonces debemos evitar que alguna de las variables ocupe toda la rejilla. Para mostrar cómo se puede dar esta situación, en la gráfica 3.3 se marcó el número de celdas para el cual el valor esperado de coincidencias entre X y Y alcanza el máximo, en está gráfica la línea roja diagonal es la identidad, nótese como para ese número de celdas la curva de celdas ocupadas por X está muy pegada a la identidad, indicando que el número esperado de celdas ocupadas por X se aproxima al total de celdas. La variable con mayor número de puntos afecta la resolución óptima en este sentido, ya que cuando la resolución es baja, además de que es más probable que existan

celdas con más de una ocurrencia, también es más probable que cubra el espacio completo, es decir, el total de las celdas. Si esto sucede la variable no nos da información sobre la otra.

El comportamiento correspondiente de ε se puede ver en las gráficas 3.1(c) y 3.1(d). En estas gráficas podemos ver que para distribuciones independientes ε es cercana a cero sin importar la resolución de la rejilla, en este sentido ε demuestra su valor, ya que sí detecta la diferencia sin importar si hay muchas o pocas coincidencias.

Cuando hay correlación, la gráfica de ε se aproxima a la forma que vimos en las gráficas de co-ocurrencias. Lo importante de esto es que nos muestra la importancia de la elección de rejilla, ya que para algunas resoluciones –demasiado gruesas o demasiado finas- ε se aproxima a cero, pero hay un rango de resoluciones en el cual detecta la asociación estadística. Además, aunque presenta ruido debido a los cambios en la agregación de puntos provocados por las fronteras de las celdas, sigue una forma que la hace buen candidato también para métodos de optimización relativamente sencillos. Esta última afirmación es cierta cuando hay dependencia, ya que, como vimos, cuando son independientes se aproxima a una constante, lo cual nos hace pensar que un método sería maximizar primero el número de coincidencias y luego refinar la búsqueda maximizando ε.

Si observamos nuevamente las gráficas de co-ocurrencias 3.1(a), 3.1(b), podemos ver que al inicio el número de co-ocurrencias crece a la par en ambas gráficas, por lo que las diferencias en ese intervalo para las gráficas de ε pueden parecer incongruentes. Lo que sucede es que N_x, el número de celdas ocupadas por X, crece más lento cuando la distribución está asociada a la distribución de Y, ya que su distribución es más densa alrededor de puntos de Y. Esto implica que ante la misma rejilla obtenemos aproximadamente el mismo número de co-ocurrencias, ya que en un principio el número está dominado por Y, tanto en el caso independiente como en el dependiente, sin embargo, el número de celdas que contienen a X es menor en el caso en que X depende de Y, es decir, hay menos celdas ocupadas por X, pero el mismo número de co-ocurrencias, lo cual resulta en una relación más fuerte entre ambas variables.

En este ejemplo, también se muestra que cuando la distribución de X depende de la distribución de Y, el número de co-ocurrencias puede alcanzar su máximo a una resolución más gruesa que la necesaria para alcanzar el máximo de ε. Esta diferencia es en parte determinada por el nivel de dependencia de X con Y. Entre más fuerte sea la correlación (si es positiva o de 'atracción') y mayor el número de puntos de X, más fina tendrá que ser la resolución, porque X estará distribuida más densamente alrededor de Y.

Hasta el momento hemos planteado implícitamente al MAUP **??** como un problema. Sin embargo, el que se obtengan diferentes resultados a diferentes resoluciones o escalas, no lo es forzosamente, en realidad, puede ser una fuente de información que nos indica el nivel al que se dan las interacciones ([45]). Poniéndolo de manera burda, si queremos estudiar bacterias con celdas de un metro, no vamos a encontrar mucho, pero al menos el resultado nos indicará que quizá estamos buscando en la escala incorrecta. Es decir, ya sea que encontremos o no una relación entre nuestras variables, hay que recordar que esto es a una resolución dada. Si optimizamos la resolución de la rejilla, la asociación estadística que obtenemos a esa resolución es sólo parte de la información, la otra parte es la resolución misma, la cual nos indica la escala a la que la señal es mayor y eso puede ser una guía igualmente útil que nos indique la escala a la cual buscar las causas por las que se da esa asociación.

Para mostrar como varía ε, en 3.4 se muestran dos mapas de calor donde se presentan los valores de ε entre el vector *L. panamensis* y los 427 mamíferos en la colecta. La información se presenta en dos mapas, en ambos los renglones corresponden a resoluciones de rejilla y las columnas a mamíferos, en el primero (3.4(a)), la coloración se generó normalizando los valores por renglón, esto nos permite ver como se comparan las magnitudes de ε a la misma resolución, también se aprecia como se acentúan las diferencias conforme se aumenta la resolución. Por ejemplo, en la resolución más baja (1 celda) no hay diferencia todos las ε son 0. En la siguiente resolución (4 celdas) aparecen las diferencias aunque las diferencias no son muy grandes. En cambio si vemos el renglón más alto

con $1,048,576$ celdas el número de asociaciones fuertes es mucho más pequeño, pero la diferencias entre los relaciones fuertes y las débiles son muy marcadas. En el segundo mapa de calor (3.4(b)) la normalización de los valores se hizo por columna, esta visualización indica como afecta el cambio de resolución en la asociación estadística por mamífero, es decir, para cada mamífero el rojo más intenso indica dónde alcanzó la asociación su valor máximo. Aquí podemos ver que hay un grupo de mamíferos que alcanzan su asociación estadística más alta en la resolución más fina (izquierda), otro grupo que en realidad no parece encontrar relaciones estadísticas a ninguna resolución (centro del mapa de calor), y otro que las encuentra en resoluciones bajas (derecha del mapa de calor).

3.4.2. Optimización de la resolución de la rejilla

Como mencionamos, un criterio que se puede aplicar es buscar un tamaño que maximice el número de celdas ocupadas por ambas variables, ya que esto es una forma de maximizar el tamaño de la muestra efectiva. El problema se puede extender para un número arbitrario de variables espaciales. Si tenemos N variables espaciales, cada pareja tendrá un máximo de coincidencias a alguna resolución, pero este será particular de cada pareja. Aunque podríamos calcular el máximo para cada pareja el resultado es confuso para el analista ya que cada asociación estadística se detecta a distinta escala. En este caso el objetivo se puede simplificar a encontrar la rejilla que maximiza la suma de coincidencias de todas las parejas de variables.

Para encontrar un tamaño que maximice el número de coincidencias, la solución directa es una búsqueda exhaustiva, definir un rango de tamaños que abarque la mayor parte de las soluciones -por ejemplo, que vaya desde un tamaño muy cercano a cero hasta otro que abarque toda la región, haciendo una subdivisión fina del intervalo de resoluciones y calculando para cada resolución las coincidencias correspondientes. El problema es que esta solución es costosa, en especial si se quiere realizar para un número grande de variables, y más aún si el número de puntos por variable es

grande. Por esta razón es útil un método de optimización que nos permita encontrar una buena aproximación a la resolución óptima.

3.4.3. Optimización del número de coincidencias

Dado un conjunto de N variables booleanas espaciales, definimos la función $C(r)$ cuyo valor es el número total de celdas con coincidencia a resolución r. Supongamos que el rectángulo mínimo que contiene todos los puntos de las distribuciones tiene celdas de tamaño $R \times R$, entonces, en particular sabemos que $C(R) = \frac{N(N-1)}{2}$. Buscamos la subdivisión del rectángulo para la cual la suma de co-ocurrencias por parejas es máximo.

1: Sean $P_1, \ldots, P_k$ una colección de distribuciones de puntos y sea $R = [x_1, x_2] \times [y_1, y_2]$, tal que $x_2 - x_1 \geq y_2 - y_1$.
2: Inicio: $r_m = \frac{x_2 - x_1}{2}$, $r_d = (x_2 - x_1)$, $r_i = s_m$, donde s_m es el tamaño más pequeño de celda.
3: **while** $r_m \neq r_i$ y $r_m \neq r_d$ **do**
4: **if** $max = C(r_m)$ **then**
5: $r_d \leftarrow r_d + \frac{r_d - r_m}{2}$ y $izq \leftarrow r_i + \frac{r_m - r_i}{2}$
6: **else if** $max = C(r_i)$ **then**
7: $r_d \leftarrow r_m$ y $r_m \leftarrow r_i + \frac{r_m - r_i}{2}$
8: **else if** $max = C(r_d)$ **then**
9: $r_i \leftarrow r_m$ y $r_m \leftarrow r_m + \frac{r_d - r_m}{2}$
10: **end if**
11: **end while**

El paso inicial es calcular el número de coincidencias para tres resoluciones, para la resolución más fina r_i, la más gruesa r_d, y para el punto medio r_m. Y actualizamos de tal forma que si el número de coincidencias $C(r_i)$ es mayor que $C(r_d)$ entonces restringimos la búsqueda al segmento izquier-

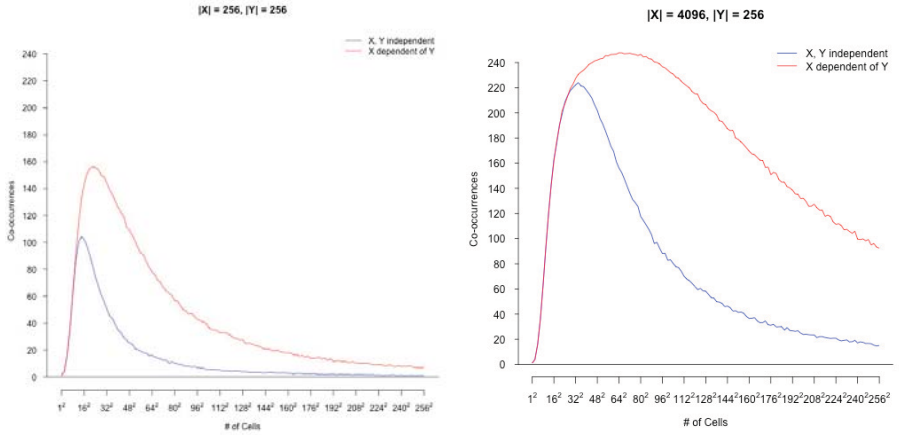
do, si es al revés restringimos al derecho y si el máximo es el punto medio entonces encogemos el intervalo a la mitad manteniendo el punto medio como tal. El algoritmo se aplicó a los datos de la base de datos de mamíferos seleccionando distintos conjuntos de variables obteniendo los siguientes resultados.

3.4.4. Resultados

En el primer experimento se utilizaron 8 especies, cuyas distribuciones suman un total de 3032 puntos. Se utilizó una búsqueda exhaustiva para un rango de tamaños de celda de 0.005 a 25 por lado, con incrementos de .005. El máximo que se obtuvo por búsqueda exhaustiva fue de 2420 coincidencias con una celda de tamaño 0.21 (3.5(a)). Por su lado el método de bisección obtuvo un máximo de 2399 con celdas de 0.12 por lado en 27 iteraciones (3.5(b)).

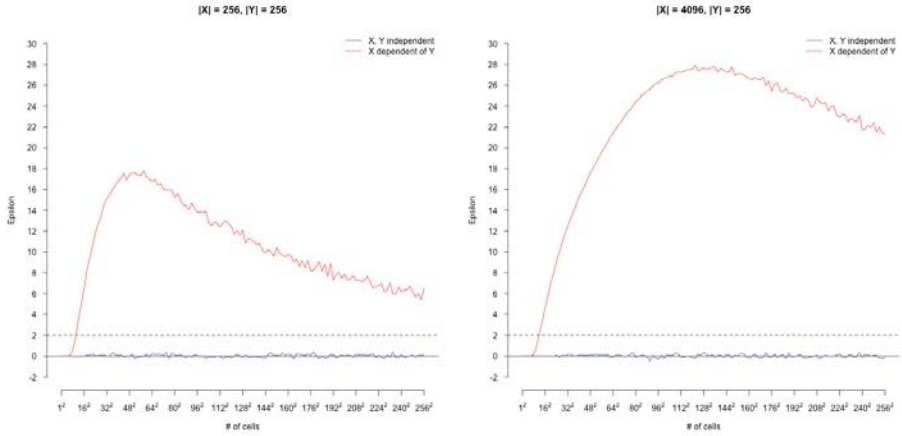
Se realizó también la optimización para el conjunto de datos completo, que tiene 438 especies con un total 35,397 puntos. En este caso el rectángulo mínimo que contiene a todos es $R = [-117.11, -86.83] \times [14.6, 32.6]$. El máximo de coincidencias se obtuvo al subdividir en 37 el rango de longitud, lo que equivale a celdas de 0.8183 por lado (3.6(a)), resultado al que llegó en 27 iteraciones y que se aproxima bastante al encontrado por una búsqueda exhaustiva con incrementos de 0.005 en el tamaño de la celda (3.6(b)).

La búsqueda exhaustiva se realizó con $5,000$ iteraciones para un rango de celdas entre 0.005 y 25 por lado, y obtuvo como máximo $307,992$ coincidencias al usar celdas de 0.82 por lado. Por otro lado, el método de bisección convergió a un tamaño de 0.8183 que da $306,979$ coincidencias. Si consideramos que para 438 variables hay $95,703$ combinaciones de dos, el máximo nos da un promedio de poco más de tres coincidencias por pareja.



(a) Variables con distribuciones de la misma cardinalidad

(b) Variables con distribuciones con cardinalidades muy distintas



(c) ε entre variable espaciales con el mismo número de instancias

(d) ε entre variable espaciales donde X tiene muchas más instancias

Figura 3.1: Co-ocurrencias y ε entre dos variables espaciales booleanas como funciones de la resolución de rejilla. La gráfica azul corresponde a variables independientes y la roja al caso donde X depende de Y.

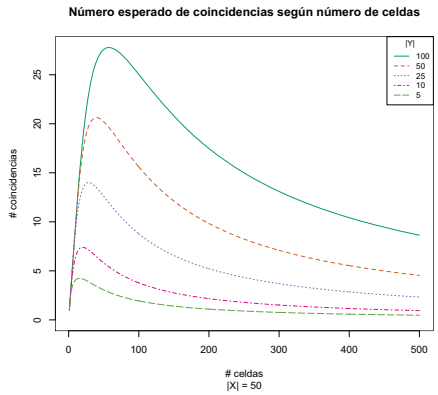


Figura 3.2: Número esperado de coincidencias entre dos variables espaciales independientes X, Y como función del número de celdas.

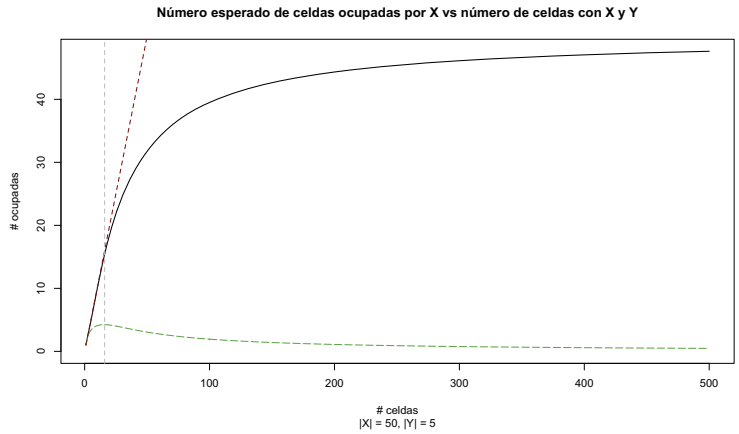
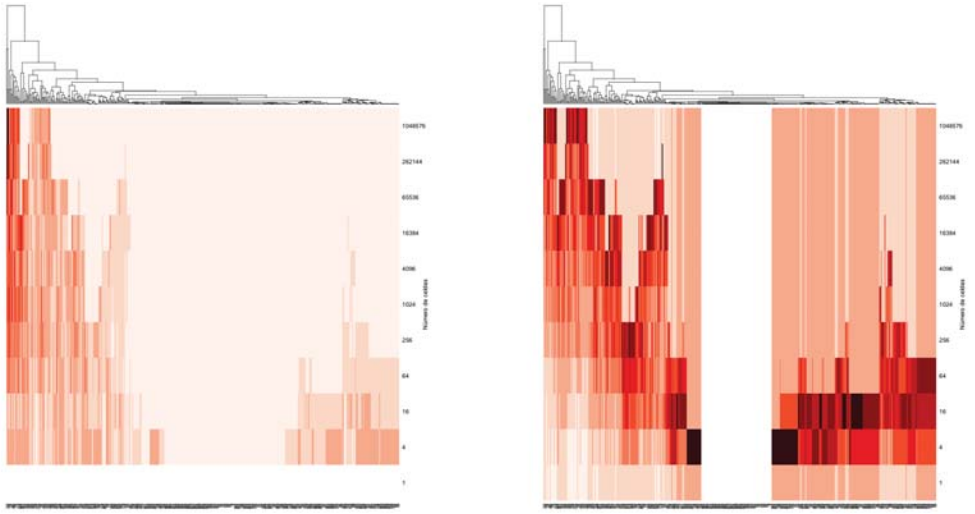


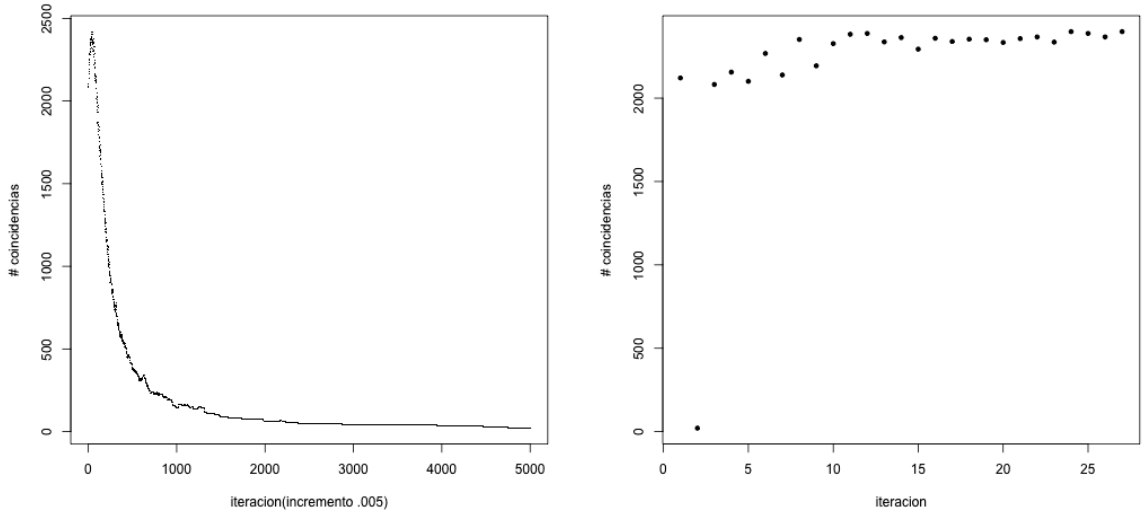
Figura 3.3: Comparación entre el número esperado de coincidencias entre X, Y y el número esperado de celdas ocupadas por X.



(a) Mapa de calor de ε para L. panamensis con los valores normalizados por renglón. Cada columna corresponde a una especie de mamífero y cada renglón a una resolución

(b) Mapa de calor de ε para L. panamensis con los valores normalizados por columna

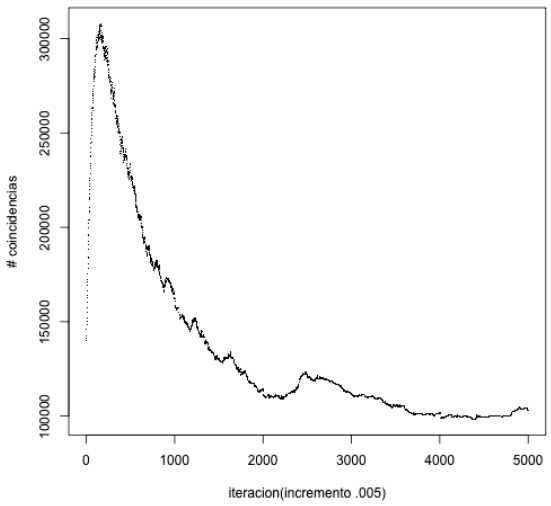
Figura 3.4: Cada columna corresponde a una especie de mamífero y cada renglón a una resolución.



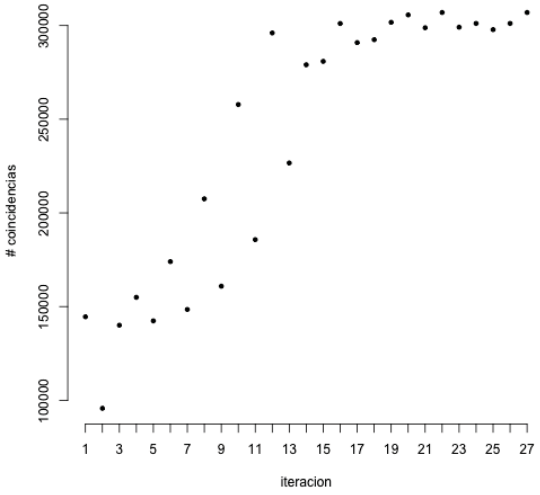
(a) Búsqueda exhaustiva

(b) Método de Bisección

Figura 3.5: Búsqueda del máximo de coincidencias para 8 especies.



(a) Búsqueda exhaustiva con 438 especies



(b) Método de Bisección con 438 especies

Figura 3.6:

Capítulo 4

Redes para el análisis de datos

Las redes (o gráficas) son por lo general la mejor estructura matemática para representar relaciones binarias entre múltiples objetos. Su origen como abstracciones de un sistema de conexiones se remonta a principios del siglo XVIII, cuando Euler resolvió el problema de "Los siete puentes de Köninsberg" utilizando una red como abstracción, donde los puentes son las aristas y los nodos eran los vecindarios que conectaban. El artículo donde Euler publicó su solución es considerado como el primer documento en teoría de las gráficas. A partir de entonces y hasta la actualidad las redes han atraído la atención de matemáticos, físicos, biólogos y sociólogos por igual. Las redes son herramientas de análisis populares porque permiten visualizar las relaciones de tal forma que es más fácil para el ojo humano detectar patrones, y porque proveen una estructura matemática muy rica que permite su análisis algorítmico. Existen varios ejemplos de aplicaciones de redes a problemas de minería de datos. Ejemplos importantes son las redes bayesianas en probabilidad [58], representación de patrones de co-ubicación espacial [61] y redes espaciales en minería de datos espaciales [47, 7, 13], redes sociales en sociología [46], redes tróficas en ecología [67, 25], y redes de co-ubicación de palabras en procesamiento de lenguaje natural [56, 10, 41].

Los términos red y gráfica se usan de forma intercambiable y depende de la disciplina la elección de

la palabra. En general, el término gráfica se usa en matemáticas, mientras que físicos, sociólogos, biólogos, y otros usan redes. Pareciera que el consenso implícito, es usar el término gráfica cuando se habla en un marco puramente abstracto, es decir, las gráficas como objetos matemáticos, y se utiliza el término red en contextos aplicados, cuando la gráfica es una representación de un sistema concreto. A lo largo de este trabajo se ha usado el término red, precisamente por esta razón, ya que el fin de este trabajo es utilizar las redes como herramienta de representación.

Una gráfica se define por un conjunto de vértices V y una relación entre ellos. Estos conjuntos determinan a la gráfica, es decir, una gráfica es una pareja de conjuntos $G = (V, E)$, donde $V \neq \emptyset$ y $E \subset V \times V$. A los elementos de V se les llama vértices o nodos, y los elementos en E aristas [33]. Si E cumple con que $\forall (a, b) \in E$, $(b, a) \in E$, entonces decimos que G es no dirigida. Las aristas también pueden tener valores asignados o pesos, por consecuencia a este tipo de gráficas se les llama pesadas, en este caso se definen como una terna $G = (V, E, w)$, donde V son los vértices, E las aristas y $w : E \to \mathbb{R}$ es la función de pesos. Este tipo de gráficas se usan, por ejemplo, en problemas de optimización para representar conjuntos de lugares y los costos para moverse de uno a otro, en la representación del sistema como gráfica los lugares son representados por los nodos, si hay una ruta entre dos lugares sin pasar por otro entonces tenemos una arista entre esos nodos, y el peso de la arista podría estar dado por el tiempo que llegar de un sitio al otro. Nuestra metodología usa redes pesadas para representar sistemas de asociaciones inferidas.

La representación visual usual de una gráfica es en dos dimensiones, donde los nodos se representan como puntos y las aristas como líneas (usualmente rectas) entre los puntos, y si la gráfica es dirigida, en vez de líneas se utilizan flechas para indicar la dirección de la relación, esta representación es la más común e intuitiva. Sin embargo, en ocasiones resulta conveniente representarla como matriz de adyacencia, ya que puede tener ventajas para detectar visualmente patrones de conexión de los nodos. En nuestro caso utilizaremos la primera, aunque no descartamos que extensiones de este trabajo se incluya la opción de usar la representación con matrices.

Por otro lado, existen distintas técnicas para dibujar una red automáticamente: Disposición circular, jerárquico (en el caso de árboles), arcos, fuerzas. Según algunos estudios, la disposición por fuerzas es la que en el caso genérico resulta más eficiente, ya que hace un buen trabajo minimizando los cruces de aristas y logra evitar el amontonamiento de los nodos dentro de lo posible. Por esta razón usamos el algoritmo de fuerzas para organizar las redes en nuestra aplicación.

La disposición por fuerzas consiste en pensar la red como un sistema de resortes, donde cada arista es un resorte y tiene una longitud de reposo y los nodos tiene una fuerza de repulsión hacia otros nodos, se definen además la fricción del sistema, y la fuerza de recuperación de los resortes. El algoritmo actualiza el sistema buscando minimizar la energía total a través de varias iteraciones hasta que se estabiliza el sistema o se llega a un número determinado da iteraciones. Existen más de una versión de esta técnica, pero todas las plataformas para análisis y visualización de redes tienen al menos una de las versiones de este método.

4.1. Aplicaciones de redes

En los párrafos anteriores se mencionó que las redes son utilizadas en distintos campos de estudio. En esta sección presentamos algunos ejemplos de dichos usos.

4.1.1. Redes espaciales

Las redes espaciales son el uso más común de redes en ciencias de la información geográfica. En realidad, la primera red en la literatura, la de los puentes de Köninsberg, es una red espacial. Las redes espaciales son redes que por lo general modelan sistemas de infraestructura, como redes de comunicaciones o redes de transporte [13], también entran dentro de este tipo redes que representan flujos entre entre localidades -por ejemplo, flujos de migración- [26]. Un ejemplo de red espacial es un sistema de carreteras representado como una red donde las aristas son las carreteras y los

nodos intersecciones entre estas, uno puede entonces utilizar medidas como el grado de centralidad de los nodos y ubicar rutas vulnerables para un sistema de manejo de emergencias [15]. En "The spatial structure of networks" [22], los autores presentan un modelo para generar redes derivado de propiedades observadas en redes espaciales, y este modelo se usa buscar a partir del modelo generador resultante por qué se presentan ciertas características que se presentan en este tipo de redes geográficas. En este artículo toman como modelos de referencia tres redes cada una de distinto origen: La red de carreteras interestatales en EEUU; la red de conexiones de la aerolínea Delta y las conexiones de datos entre sistemas autónomos de internet en EEUU. En el artículo muestran que la red de carreteras es casi plana (es decir se puede dibujar casi sin que las aristas se intersequen), mientras que las otras dos no. A partir de esta observación construyen un modelo para generar redes basado en optimización del costo de la red de acuerdo al costo de la conexión entre cada par de vértices, costo que depende de dos componentes: la longitud de las aristas y el número de vértices que hay que recorrer para llegar de un lado a otro. La importancia de cada uno de estos componentes en el costo es dada por un parámetro que representa las preferencias del usuario. Por ejemplo, en vuelos la preferencia del son viajes con el menor número de paradas posibles, mientras que para un viaje en carretera el usuario busca minimizar el número de kilómetros. Los autores muestran que utilizando este criterio si se le da mayor peso al número de vértices en el costo de un camino se generan redes parecidas al de la red de aeropuertos y la de internet, y si se le da todo el peso a la longitud las aristas se genera una red parecida a la red de carreteras en el sentido de que es una red casi plana. Lo interesante es que con un modelo bastante simple (sólo depende de un parámetro) describen cualitativamente una familia de redes espaciales y muestran un camino para ir desarrollando una teoría sobre redes espaciales. Otro de los puntos importantes que se mencionan en este artículo es que el componente geográfico en muchos casos no es incluido en el análisis de redes espaciales y este estudio muestra además la importancia de considerar ese componente ya que su modelo se deriva de observaciones sobre propiedades geográficas, en este caso principalmente

distancias.

4.1.2. Modelos gráficos

En el caso de modelos gráficos, como las redes bayesianas, la red representa un conjunto de variables aleatorias y sus dependencias condicionales vía una gráfica acíclica dirigida. Por ejemplo, una red bayesiana, podría representar las relaciones entre síntomas y enfermedades. Dados ciertos síntomas uno podría calcular la probabilidad de que sean producidos por enfermedades particulares. Formalmente, las redes bayesianas son redes acíclicas dirigidas donde los nodos representan variables aleatorias y las aristas dependencias condicionales entre las variables, de tal forma, que para cualquier nodo N tenemos que el nodo es independiente de la red dado el conjunto de nodos que apuntan a éste. Por ejemplo, en 4.1, cada nodo está asociado con una función de probabilidad de tal forma que para cada combinación de valores de verdad de los padres nos da la probabilidad de que éste suceda. Además nos da las relaciones de independencia condicional entre las variables, por ejemplo en este caso, D es independiente de A dados B y C. En el caso general, dado un nodo, si este tiene m padres, su función de probabilidad condicional se representaría por una tabla de $(m+1)\times 2^m$, esto es un renglón por cada combinación de valores de verdad de los m padres. Los modelos gráficos representan las relaciones algebraicas entre variables aleatorias, como la regla de la cadena e independencia condicional, y su uso más común es para hacer inferencias probabilísticas como calcular la probabilidad a posteriori.

4.1.3. Redes en procesamiento del lenguaje natural

Las redes de co-ocurrencia de palabras son utilizadas en procesamiento del lenguaje natural, y, a diferencia de las redes espaciales estas son redes inferidas a partir de la distribución de palabras en cuerpos de texto. Este tipo de metodología es similar a la metodología que se desarrolla en seccio-

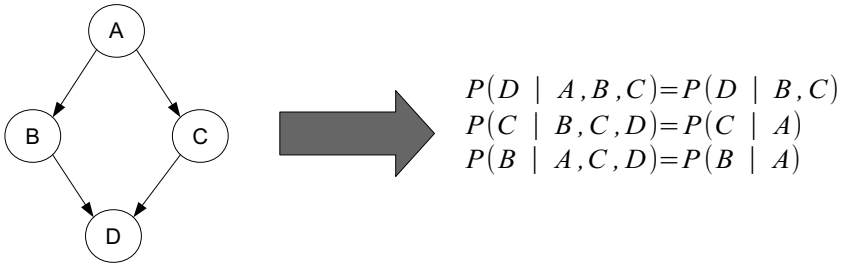


Figura 4.1: Ejemplo de red bayesiana de cuatro variables, esta red implica por ejemplo $P(A,B,C,D) = P(A)P(B \mid A)P(C \mid A)P(D \mid B,C)$

nes posteriores de este trabajo. En procesamiento de lenguaje natural las redes son construidas en escenarios como descubrimiento de tópicos[[40]], o para representar un texto y visualizar el historial de cambios que ha sufrido[[41]]. Este tipo de redes, donde los nodos representan palabras y existe una arista entre dos nodos si las palabras se encuentran cerca frecuentemente, son conocidas como redes de co-ocurrencia de palabras. Para construir este tipo de redes se define una distancia máxima entre palabras para que se considere una co-ocurrencia, por ejemplo si se define 2 como la distancia máxima entonces para dos palabras dadas, por cada vez que aparezcan en un texto a dos o menos palabras de distancia se considerará que tienen una co-ocurrencia. Existen distintas versiones de esta definición en algunos casos además de la distancia, se pide que formen parte de la misma oración, en otros casos no se pone esa restricción o se pueden considerar distancias mucho más grandes, por ejemplo de 50 palabras, obviamente, todo depende de los objetivos con que se

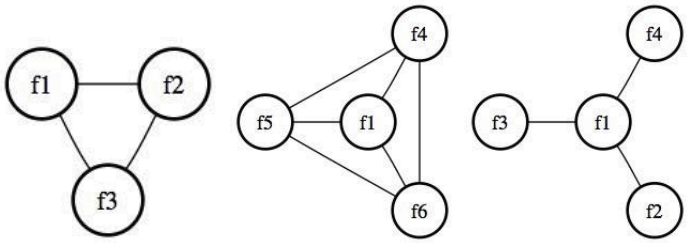
construye la red. Una vez definidas las reglas de co-ocurrencia para cada par de palabras se evalua si estas co-ocurren con mayor frecuencia de lo esperados, por ejemplo, si para dos palabras W_1 y W_2 tenemos que $P(W_1, W_2) < P(W_1)P(W_2)$ [[10]]. En procesamiento de lenguaje natural o lingüística de corpus, las redes son usadas en distintos temas como redes de adyacencia de palabras, redes semánticas, redes de co-asociación y redes sintácticas [[3]]. En "Conceptual grouping in word co-occurrence networks"[56], se construyen redes de co-ocurrencia de palabras para un sistema de recuperación de información. En ese trabajo dos palabras co-ocurren si están a menos de 50 palabras de distancia en el mismo texto, y la medida de la intensidad de su asociación se calcula a partir de la frecuencia con que esto ocurre en una base de datos de textos. En "Choosing the word most typical in context using a lexical co-occurrence network" [[17]] se construye una red de palabras usando t-scores para cuantificar la significatividad de las asociaciones. En "Visualizing Sequences of Texts Using Collocational Networks" [[41]], las redes de co-ocurrencia de palabras son utilizadas para caracterizar una serie de reportes técnicos y visualizar de manera concisa las diferencias entre estos. Los ejemplos anteriores son muestras del uso de redes para caracterizar documentos. También son ejemplos de como una red puede ser usada para extraer información algorítmicamente usando su topología. Por ejemplo: para demostrar que el lenguaje tiene una estructura de mundo pequeño [[10]]; para seleccionar automáticamente el significado correcto de una palabra según su contexto [[17]]; o como puede ser usada como una herramienta de visualización [[41]].

4.1.4. Patrones de co-ubicación

Los patrones de co-ubicación son otra manifestación de redes que aparece en minería de datos espaciales. Como se mencionó en el capítulo 3. Dado un conjunto F de variables o características booleanas espaciales, un patrón espacial P es un subconjunto de esas variables, $P \subseteq F$.

Un patrón de co-ubicación espacial puede ser representado como una red no dirigida, donde un

nodo corresponde a una variable espacial booleana y una arista a un a una relación de co-ubicación entre las variables. Al parece existen dos formas de definir una co-ubicación en la literatura: una es definirla como un subconjunto de variables cuyas instancias forman un clique usando una relacion R [66] (Fig. 4.2(a) y 4.2(b)); una definición más relajada es en la que la no hay el requisito de que sea un clique, es decir, puede no haber una arista entre dos variables del patrón [68] (Fig. 4.2(c)).



(a) Clique de tres variables (b) Clique de 4 variables (c) Definición relajada no requiere que sea un clique.

Figura 4.2: Tipos de red que pueden definir patres de co-ubicación.

Las redes ofrecen una representación visual y matemática que puede ser usada en algoritmos de minería, pero parecen ser usados únicamente en la segunda forma abstracta. Parece haber algún trabajo en proceso para usar la representación visual que proveen las redes de co-ubicación para estudiar la distribución geográfica de los patrones de co-ubicación [42], pero hasta donde sabemos no hay mucho trabajo en esa dirección.

Lo que quisiéramos recalcar es que el objetivo de un patrón de co-ubicación es explicar un fenómeno a una escala local. Esto es un patrón de co-ubicación representa un patrón local que aparece frecuentemente, y en ese sentido, las redes asociadas están aisladas las unas de las otras. Por ejemplo, supongamos que hay cuatro variables espaciales booleanas f_1, f_2, f_3, f_4, con una distribución tal que para la variable f_1, tenemos los patrones de co-ubicación $C_1 = f_1, f_2, f_3$ y $C_2 = f_1, f_4$. Lo que las redes no muestran explicitamente es que f_2 y f_3 pueden tener un efecto indirecto en la distribución

de f_4.

4.2. Redes y análisis de datos espaciales

La minería de datos está compuesta esencialmente por cinco tareas: preparación y limpieza de datos; análisis exploratorio de datos; generación de modelos descriptivos; descubrimiento de patrones y generación de modelos de predicción [29]. Todas estas tareas se complican mucho por la 'maldición de la dimensionalidad', cuando el número de variables es grande. Existen varios métodos para lidiar con éste problema, uno de ellos es el análisis exploratorio de datos para implementar selección de características [29]. De tal modo que sólo un subconjunto de variables son utilizadas en el resto del proceso de minería de datos. En minería de datos espaciales la maldición de la dimensionalidad y el problema de la selección de características es aún más agudo. ¿Cómo escoge uno de entre las distribuciones espaciales de muchas variables cuáles son importantes para predecir la distribución de otra variable espacial? Una red que es simplemente la representación de relaciones conocidas no nos es de utilidad, ya que ni siquiera conocemos *a priori* que relaciones existen de entre una gran cantidad de posibilidades y mucho menos sabemos cuáles son importantes.

Entonces, la pregunta es si las redes no serían útiles de manera más general desde la perspectiva de la minería de datos. Dado que las redes son una forma poderosa de representar relaciones entre objetos, es deseable una metodología que nos permita tenerlas disponibles como parte de nuestra caja de herramientas de análisis. Por supuesto, esto implica por lo general que las relaciones sean inferencias estadísticas –en lugar de relaciones conocidas– cuyo grado de certeza es cuantificado relativo a una hipótesis nula–por ejemplo, que la variable es independiente. Estas relaciones inferidas pueden ser usadas después para la construcción de teorías o modelos que pueden ser independientemente validados. A continuación se presenta un método de inferencia de asociaciones entre características espaciales booleanas, para luego representar el sistema de relaciones

y características con redes que permitan explorar de manera eficiente el espacio de asociaciones estadísticas usando una visualización interactiva de la o las redes.

El método presenta similitudes con el proceso de descubrimiento de patrones de co-ubicación[3.3.1]. Sin embargo, existen diferencias fundamentales en los objetivos y los resultados que se obtienen. Aunque en ambos se infieren las asociaciones entre variables a partir de sus distribuciones espaciales, los patrones de co-ubicación describen subconjuntos de variables que ocurren frecuentemente, pero estos patrones no están conectados con los demás patrones de co-ubicación, al menos explícitamente. En otras palabras los patrones son vistos como partes aisladas. En contraste, el modelo que se construye con nuestra metodología está basado en un análisis global, de tal forma que la red da una vista del sistema multivariado como un todo, y se pueden encontrar conexiones entre conjuntos de variables que en otro tipo de modelos parecen desconectadas.

Una característica de esta metodología es que las redes que produce son redes pesadas y dirigidas, y por lo tanto ofrecen información sobre la intensidad de una relación y su significado. Los pesos además son importantes como herramienta de exploración, ya que pueden ser usados para definir filtros que eliminan aristas cuyo peso esta por debajo o por arriba de cierto umbral. Definir este tipo de filtros permite analizar el sistema desde distintas perspectivas y desde distintos niveles de complejidad. Los pesos no sólo representan la intensidad de una relación (magnitud), también indican el tipo de relación en términos de ‘atracción’ (valores positivos/correlación positiva) y ‘repulsión’ (valores negativos/correlación negativa). Dado que las aristas son construidas a partir de las co-distribuciones espaciales, si hay relaciones que se derivan de una región geográfica, entonces pueden existir cúmulos de nodos que representan una regionalización de la red. Por ejemplo, características que están relacionadas porque son endémicas de una misma región, formarán subredes densamente conectadas.

Además del filtro de aristas por pesos, existen otros dos filtros en el proceso. Uno se aplica en el momento en que se seleccionan los tipos de nodos, esto es, cada característica espacial booleana

puede tener asignada una categoría o tipo, como enfermedades y síntomas, o mamíferos e insectos. Cuando la red es visualizada se podría elegir sólo ver un subconjunto de las categorías disponibles. Esto constituye un filtro de tipos de nodo. El otro tipo de filtro es cuando uno decide cuales son las relaciones de interés. Si, una vez más, consideramos categorías de nodos, podemos elegir de entre tener todos contra todos, es decir, si tenemos categorías $C_1 \times C_2 \times \cdots \times C_n$ el conjunto de relaciones $C_i \times C_j$, donde $i, j \in \{1, \ldots, n\}$, o podríamos restringir el análisis a pares en $C_1 \times C_2$. Más adelante, en 4.3.2, presentamos un caso de estudio donde se aplican estas ideas.

Como nota al margen, cabe mencionar que aunque en este trabajo se consideran únicamente relaciones binarias, se podrían considerar relaciones de mayor cardinalidad. Podríamos definir, por ejemplo, relaciones ternarias en las que se representa para cada tres variables X, Y y Z la probabilidad $P(X, Y, Z)$. En este caso estamos pasando de redes a hipergráficas, es decir gráficas donde una arista une más de tres nodos. El problema es que las hipergráficas son mucho más difíciles de visualizar de manera efectiva, y su visualización no está tan bien entendida [34], especialmente si hablamos de hipergráficas grandes con relaciones dirigidas.

Es importante señalar que existe una vasta bibliografía sobre el análisis de redes en términos de cuantificar y resumir su estructura y sus propiedades [24]. Por ejemplo, conectividad de nodos, análisis de cúmulos y estructura comunitaria, así como medidas de similitud entre nodos, detección de módulos o caminos de relaciones de alta significatividad usando los pesos de aristas. Hay muchas medidas y algoritmos que pueden ser usados para hacer minería de datos en redes [11, 39]. Ejemplos de esto son el indice de ley de potencias, algoritmos de detección de comunidades [44, 43], o algoritmos de clasificación de nodos más generales [36]. El desarrollo de técnicas de visualización de redes son también una parte activa de este campo, con resultados como substratos semánticos para la disposición de nodos[4], visualización simultánea de conectividad local de los nodos mediante su matriz de adyacencia [31] o técnicas para el etiquetado automático nodos [16]. Además existe una rama dedicada a definir formas de evaluación de las técnicas de visualización [32]. Varias

de las técnicas mencionadas han sido empleadas en el análisis de redes espaciales [22, 13, 27, 60]. En resumen, hay una plétora de propiedades que pueden descubrirse acerca de un sistema mediante el análisis visual y algorítmico de sus redes inferidas.

4.3. Redes inferidas para datos espaciales

Esta metodología construye redes a partir de datos espaciales, donde las aristas representan asociaciones estadísticas entre variables, el método a través del cual se construyen se compone de cuatro pasos. El primero es definir un modelo de co-ocuerrencia local para relacionar características espaciales booleanas entre sí 3.3.1. El siguiente, definir el tipo de asociaciones en que estamos interesados, es decir, aquí especificamos si nos interesa cualquier relación entre dos variables o sólo las relaciones entre ciertas clases de variable. El tercero, calcular las estadísticas de diagnóstico a partir de las co-ocurrencias que pertenecen al tipo de relaciones en que estamos interesados. Finalmente, se construye la red para visualizar la estructura, donde los nodos representan a las variables, y las aristas asociaciones significativas entre ellas.

Por simplicidad, usamos un modelo de co-ocurrencia basado en una rejilla uniforme, de tal forma que si un conjunto de características tienen instancias en una misma celda entonces se define que tienen una co-ocurrencia. Existen otros modelos que podrían usarse para el mismo fin, como definir un radio de influencia alrededor de cada punto. Pero, el método que elegimos requiere menos recursos de computo y se puede implementar fácilmente en sistemas de bases de datos geográficos como PostGIS.

El segundo paso es especificar el tipo de asociación. Un ejemplo es simplemente la co-ocurrencia, $C_{1,2,\ldots,n}$, de un conjunto de variables o características $X_1, X_2, \ldots, X_n$. Reglas más generales se pueden construir, como es usualmente el caso en patrones de co-ubicación. Por ejemplo, en lugar de asociaciones ligadas por el ‘y’ lógico, como en $X_1 \wedge X_2 \wedge X_3$ co-ocurren, uno podría usar el ‘o’ lógico, por

ejemplo $X_1 \wedge X_2 \vee X_3$ co-ocurren. Una vez que el tipo de asociaciones está definido podemos inferir las dependencias estadísticas entre características espaciales booleanas, como la dependencia de la distribución de una variable X_i en la distribución de otra variable X_j. Esta dependencia estadística puede ser a su vez definida de varias formas. Una opción natural es $P(X_1, \ldots, X_n)$. Si $N_{1,\ldots,n}$ es el número de celdas que contienen co-ocurrencias de $X_1, X_2, \ldots, X_n$, y N es el número de celdas en la rejilla, entonces

$$\hat{P}(X_1, X_2, \ldots, X_n) = \frac{N_{1,2,\ldots,n}}{N} \tag{4.1}$$

estima la probabilidad $P(X_1, X_2, \ldots, X_n)$ de que las características co-ocurran. También es natural considerar probabilidades condicionales como $P(X_i \mid X_j)$, las cuales pueden ser aproximadas por

$$\hat{P}(X_i \mid X_j) = \frac{N_{i,j}}{N_j} \tag{4.2}$$

Una pregunta interesante es qué tanto la frecuencia con que se da un patrón de co-ocurrencia dado se aleja de lo que uno esperaría. Por ejemplo, $P(X_i \mid X_j) - P(X_i)$ cuantifica la diferencia con que X_i ocurre en presencia de X_j en relación con su distribución independiente de X_j, en otra palabras, es una forma de medir si X_j nos da alguna información adicional sobre la distribución de X_i. En este caso, $P(X_i)$ sirve como hipótesis nula con respecto a $P(X_i \mid X_j)$. Con el objetivo de incluir una medida de confianza, el problema puede ser planteado directamente como una prueba de hipótesis. Formalmente, definimos dos variables aleatorias S_i y $S_{i|j}$, donde S_i representa el número de ocurrencias de un evento con probabilidad $P(X_i)$ de ocurrir y $S_{i|j}$ el número esperado de éxitos de un evento con probabilidad $P(X_i \mid X_j)$ de ocurrir, ambos en N_j intentos. Esto define dos variables aleatorias con distribución binomial que modelan el comportamiento de X_i en el caso independiente y en el caso vinculado a X_j.

El valor esperado de una variable binomial generada en n intentos por un evento con probabilidad

p de ocurrir es np y su varianza está dada por $np(1-p)$. Por lo tanto, para S_i y $S_{i|j}$, tenemos que sus valores esperados pueden ser aproximados por

$$E(S_i) = N_j \hat{P}(X_i) \tag{4.3}$$

$$E(S_{i|j}) = N_j \hat{P}(X_i \mid X_j) \tag{4.4}$$

y la desviación estándar de S_i por

$$\sigma_{S_i} = \sqrt{N_j \hat{P}(X_i)(1 - \hat{P}(X_i))} \tag{4.5}$$

Con lo cual podemos medir la influencia de X_j en la distribución de X_i con una prueba binomial, a la cual denominamos ε:

$$\varepsilon_{i|j} = \frac{E(S_{i|j}) - E(S_i)}{\sigma_{S_i}} \tag{4.6}$$

$$= \frac{N_j(\hat{P}(X_i|X_j) - \hat{P}(X_i))}{\sqrt{N_j \hat{P}(S_i)(1-\hat{P}(S_i))}} \tag{4.7}$$

La magnitud de esta medida indica el grado al cual la hipótesis nula es violada. Salvo en casos donde las muestras son muy pequeñas, una distribución binomial puede ser aproximada por una distribución normal. De tal manera que $|\varepsilon_{i|j}| \geq 2$ correspondería a que el número de ocurrencias en presencia de X_j está fuera del 95 % de los casos si X_i fuera independiente. Otro elemento es el signo de ε, el cual determina si la hipótesis nula es violada porque el resultado es menor de lo esperado o mayor. En el caso en que $\varepsilon_{i|j} \geq 2$ quiere decir que, en presencia de X_j, X_i ocurre con significativamente más frecuencia de lo que uno esperaría si las variables fueran independientes, se podría decir que confirma que los eventos se "atraen". Por el contrario si se tiene $\varepsilon_{i|j} \leq -2$, X_i y X_j

co-ocurren con mucho menor frecuencia de lo que uno esperaría si fueran independientes. En este caso los eventos se "repelen".

En el caso de pares de variables, el número de asociaciones estadísticas posibles tiene crecimiento de orden cuadrático. Para n variables, hay $n(n-1)$ relaciones posibles. Sin embargo, el número de relaciones significativas –relaciones para las cuales $|\varepsilon| \geq t$, para algún umbral t– puede ser mucho menor. Un ejemplo se puede ver más adelante en el caso de estudio.

Hasta ahora hemos revisado los fundamentos estadísticos para una medida que, al estar asociada a estadísticas de diagnóstico, infiere interacciones probables entre características espaciales booleanas. Aunque en este caso utilizamos una prueba de hipótesis binomial estándar para medir el grado de asociación estadística, se pueden utilizar otras estadísticas de diagnóstico. Nuestro objetivo ahora es explorar estas posibles interacciones para entender no sólo qué características se afectan directamente, sino también como el efecto de una se transmite a través de una cadena de interacciones y a través del espacio, y explorar como se conecta todo. Como habíamos anticipado, una red es una representación natural que nos puede ayudar en el análisis de un sistema y ε es una herramienta que nos permite construir dicha red.

La red se define de manera natural: cada nodo N representa una característica espacial y para cada par de nodos N_i, N_j, existe una arista si $\varepsilon_{i|j}$ cumple con las restricciones especificadas, además el peso de la artista es precisamente $\varepsilon_{i|j}$. La red es, por lo tanto, una herramienta para explorar visualmente el sistema completo de asociaciones potenciales, lo que nos permite observar su complejidad, al mismo tiempo que nos brinda una estructura para combinar el análisis exploratorio con el algorítmico.

En conclusión, para determinar la relación entre variables espaciales se nos presenta un primer reto ya que debemos definir un marco de referencia para determinar cuándo las ocurrencias de dos o más variables coinciden por su ubicación. Este problema no es particular del espacio geográfico. En realidad es un problema que se presenta cuando nuestro espacio de referencia es continuo.

Llamamos “espacio de referencia” al espacio en el que se distribuyen las variables, por ejemplo, el tiempo o el espacio geográfico.

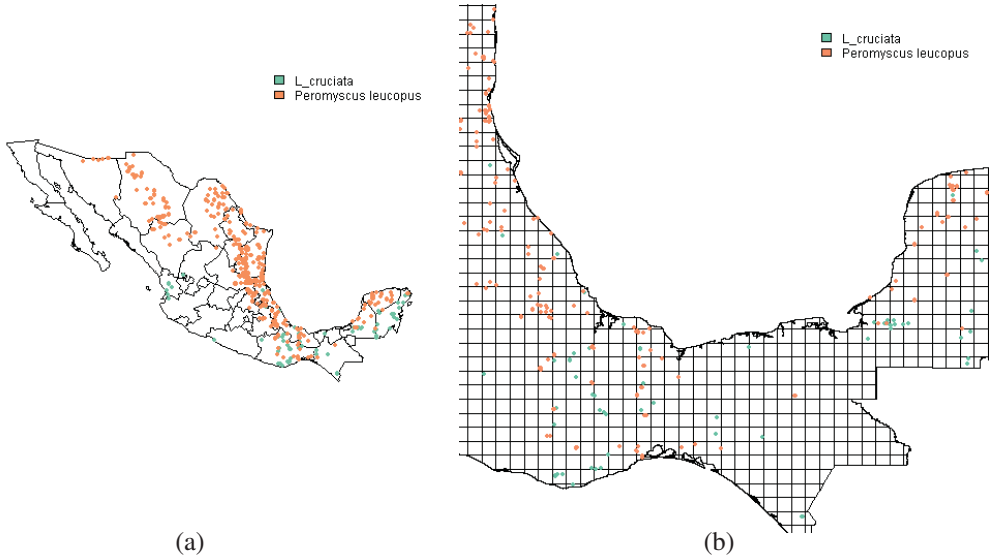
4.3.1. Caso de estudio

Para ilustrar la metodología, en esta sección consideramos una aplicación del método al estudio de la propagación de una enfermedad en México a partir de las interacciones entre varias especies animales [25]. Este es un proyecto de investigación interdisciplinario, donde especialistas de varias disciplinas están estudiando la fenomenología de enfermedades infecciosas transmitidas por vectores como la leishmaniasis. Una enfermedad como esta depende de interacciones bióticas entre especies que sirven como vectores y especies reservorio. En epidemiología, los vectores son especies, usualmente insectos, que sirven como conductos para la transmisión de agentes infecciosos, y los reservorios son especies que alojan el patógeno. La leishmaniasis es una enfermedad causada por una parásito que transmiten especies del género *Lutzomyia* (vectores) –un género de mosquito– a mamíferos que sirven como reservorio.

Una *Lutzomyia* adquiere el parásito después de alimentarse de un mamífero infectado (reservorio). Subsecuentemente, la *Lutzomyia* puede infectar otros mamíferos; en particular puede picar e infectar seres humanos. Para entender el ecosistema que da soporte a esta enfermedad y prevenir su contagio, es necesario entender las interacciones entre las especies de *Lutzomyia* y los reservorios. Sin embargo, no existe suficiente información sobre que mamíferos actúan como reservorios, y realizar estudios exhaustivos es prohibitívamente costoso en tiempo y dinero. El enfoque de este estudio fue inferir reservorios potenciales a partir de datos de colectas y estudiar las interacciones entre especies potenciales mediante la red ε resultante.

El conjunto de datos consiste de 438 distribuciones geográficas de especies en México –recopiladas a lo largo de 100 años–, las cuales suman en total 35, 397 puntos. Dado que el estudio está enfocado

a descubrir qué mamíferos pueden estar actuando como fuente de alimento para las *Lutzomyia*, estamos sólo interesados en las relaciones entre *Lutzomyia* y especies de mamífero. Más aún estamos interesados en explorar con qué especies las *Lutzomyia* tienen una correlación positiva. Lo primero por lo tanto es separar los conjuntos de puntos en dos categorías: 427 especies de mamífero y 11 especies de *Lutzomyia*. Esto quiere decir que tenemos 11 características espaciales booleanas y queremos entender como son afectadas por las otras 427 características espaciales booleanas. Es claro que explorar visualmente las 11 co-distribuciones espaciales no es una opción. La imagen 4.3(a) presenta un ejemplo con las distribuciones de dos especies: una *Lutzomyia* y un mamífero y como se ve su proyección en la rejilla 4.3(b).



(a) (b)

Figura 4.3: Distribuciones geográficas de *Lutzomyia* cruciata y Peromyscus leucopus.

Dados los objetivos de esta investigación, las relaciones entre las especies son filtradas de dos formas. Primero, en lugar de explorar todas las posibles relaciones por parejas, el análisis sólo considera el las relaciones definidas entre pares ordenados (L_i, M_j), donde L_i es una especie de *Lutzomyia* y M_j una especie de mamífero, porque como se explico anteriormente, lo que se busca

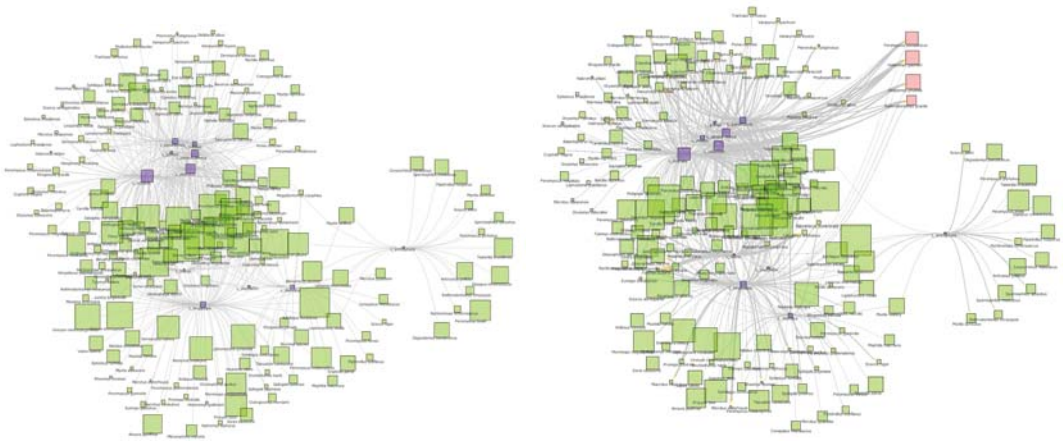
es descubrir mamíferos de los que las *Lutzomyia* puedan estar dependiendo. En segundo lugar, definimos dos como valor mínimo de ε para que la relación se considere significativa –se podrían considerar también relaciones para las que $\varepsilon \leq -2$, pero en este caso no es interesante conocer donde hay poca probabilidad de que haya presencia de *Lutzomyia*.

Después de aplicar el proceso descrito en la sección anterior, usando una rejilla de 25km de resolución[4.3(b)], encontramos 521 relaciones estadísticamente significativas, de las 4679 posibles, lo cual implica una reducción importante sobretodo si recordamos que el objetivo es permitir al investigador que explore las relaciones. El número de posibles reservorios también se redujo y de los 427, únicamente 187, tenían al menos una asociación significativa con alguna *Lutzomyia*. Lo cual muestra las aplicaciones de ε para selección de características.

4.3.2. Aplicación

Para este estudio se desarrolló una aplicación básica que construye una red inferida a partir de distribuciones de características espaciales que han sido previamente divididas y etiquetadas en dos o más categorías –en el caso de leishmaniasis las categorías vector y mamífero–. Dadas estas categorías la aplicación permite al usuario definir que relaciones son de interés –en este caso se usaron las relaciones mamífero → vector. Cabe aclarar que se pueden usar más de dos categorías –por ejemplo si además de las anteriores se tuviera la categoría reptil, se podrían definir las relaciones reptil → vector–, la aplicación simplemente espera un conjunto de etiquetas para identificar las características objetivo y otro conjunto de etiquetas que indica las características que se espera usar como predictivas. Una vez construida la red, se presenta al usuario una visualización interactiva de la red que tiene las siguientes herramientas: zoom y paneo; dos controles deslizables–uno para valores positivos y otro para valores negativos– que filtran aristas según umbrales de peso; y un campo de búsqueda que permite resaltar nodos según sus nombres (p.ej. nodos resaltados en la

imagen 4.4(b)). Otra característica de la aplicación es que usa un algoritmo de disposición de nodos dirigido por fuerzas para acomodar los nodos. El usuario puede detener el algoritmo de despliegue automático y acomodar manualmente los nodos a su conveniencia (p.ej los nodos resaltados en 4.4(b) fueron colocados manualmente).



(a) Red inferida con $\varepsilon \geq 2$

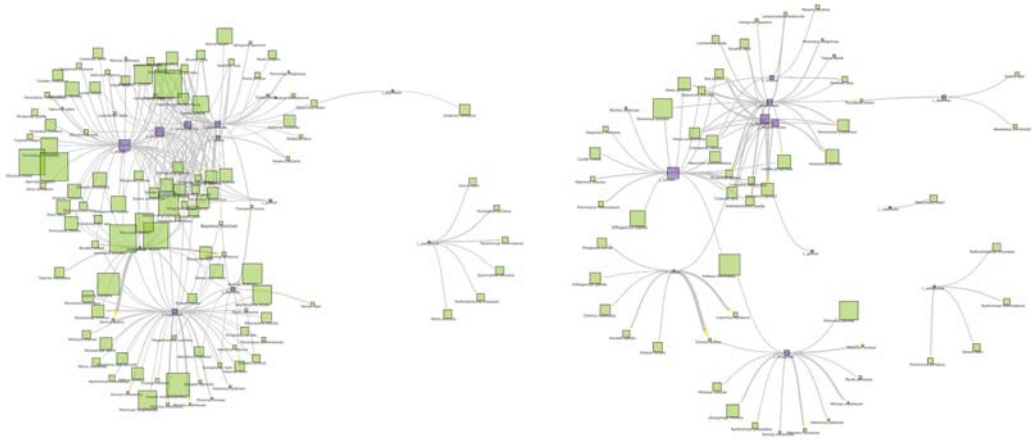
(b) Cuatro reservorios confirmados muestran asociaciones significativas con siete *Lutzomyia*

Figura 4.4: Red vector–reservorio.

Nótese que la red resultante además de ser pesada es dirigida debido a que ε no es una relación simétrica, ya que aunque $\varepsilon_{i|j}$ y $\varepsilon_{j|i}$ tienen el mismo signo no necesariamente tienen la misma magnitud. Para la visualización de la gráfica, los nodos se distribuyeron con un algoritmo de distribución dirigido por fuerzas, en el cual se busca el punto de estabilidad de un sistema donde los nodos son considerados como partículas que se repelen y las aristas como resortes que sostienen a los nodos y que tienen una misma longitud de reposo. Dicho algoritmo se usa comúnmente para generar disposiciones automáticas de los nodos de una red. En este estudio, a partir del acomodo automático de nodos se pueden identificar tres áreas de la red, cada una distribuida alrededor de un grupo de *Lutzomyia* diferente. A la derecha tenemos un nodo vector (*L. anthophora*) rodeado por un conjunto de nodos de mamíferos, la mayoría de estos sólo conectados a esta especie. Otra

sección de la red, abajo a la izquierda, se compone de cuatro especies de *Lutzomyia* y sus vecinos correspondientes, y, finalmente la tercer área está arriba, con seis vectores, y un conjunto grande de mamíferos que los rodean. Estas áreas, como veremos más adelante, corresponden de manera cercana con las distribuciones geográficas de las especies.

Aunque consideremos que dos es un umbral adecuado, uno puede tomarlo como punto de partida e ir probando otros límites. En esta aplicación el usuario puede experimentar con esto usando los controles deslizables. Esto ayuda a simplificar la red y explorar las relaciones a distintos niveles de magnitud. Las imágenes 4.5(a) y 4.5(b) muestran los efectos de mover el límite a valores más altos de ε de tal modo que sólo las asociaciones positivas más significativas se muestran. Esto resulta en redes más simples, pero que en mantienen aproximadamente la estructura de la anterior. Esta metodología puede ser usada también en selección de características, esto es, en la selección de conjuntos más pequeños de características para continuar el análisis y distinguir que características de sus distribuciones espaciales están determinando la conectividad de la red.



(a) Los efectos de elegir un límite mayor para ε (b) Resultados de eliminar aristas que no cumplen $\varepsilon \geq 4$

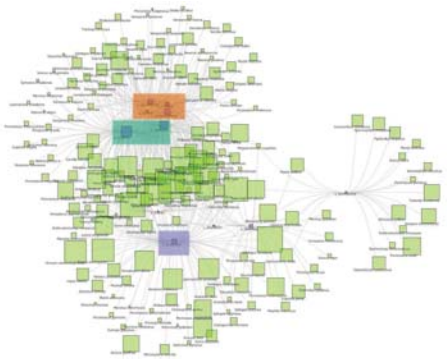
Figura 4.5: Resultado de eliminar aristas que no cumplen $\varepsilon \geq 6$.

Es importante notar que la topología de la red está en correspondencia con las distribuciones

geográficas de las especies (imagen 4.6). Por ejemplo, los nodos superiores (rectángulo naranja) corresponden con las especies *L. panamensis*, *L. ovallesi*, *L. olmeca olmeca*, *L. gomezi*, todas especies que se encuentran principalmente al sureste de la península de Yucatán. Un poco más abajo en la red, pero aproximadamente en la misma área (rectángulo verde), encontramos los nodos de *L. cruciata* y *L. shannoni* que también se encuentran en la península pero se extienden más hacia el centro de la república. En esa zona comparten espacio con *L. longipalpis* (rectángulo morado), lo cual se refleja en la red por un cúmulo de nodos compartidos. En el extremo derecho de la red, tenemos el nodo de *L. arthophora* una especie que se encuentra principalmente al norte de México y en Estados Unidos. La distribución de este vector no se traslapa con la distribución de las demás especies de vector, lo cual se refleja en una conectividad más débil de este nodo y sus vecinos con el resto de la red. Como acabamos de ver, en este caso, la estructura de la red está íntimamente relacionada con una regionalización del conjunto de datos. Además, dado que la red es conexa, esto podría indicar formas en que la enfermedad migra de una región a otra.

La red puede codificar más información que sólo las asociaciones entre especies. En esta visualización, los nodos y las aristas contienen tres piezas de información extra en sus atributos visuales. Por el color, podemos ver si un nodo pertenece a un vector o a un mamífero, el área del nodo se relaciona directamente con el tamaño relativo de la muestra correspondiente a la especie y el grosor de una arista está directamente relacionado con el grado de confianza en la relación. Por ejemplo, en este caso la combinación de tamaño de nodos con grosor de aristas puede indicarnos el tipo de relación que se está observando, una arista gruesa conectando dos nodos pequeños representa una asociación fuerte entre dos eventos raros –especies poco comunes en este contexto–[4.7].

Existen otras cosas que se pueden observar a partir de la estructura de la red. Como ya mencionamos, que la red sea conexa puede implicar que para cualesquiera dos especies, existe un camino por el cual el parásito se puede propagar de una a la otra. El grado de conectividad es, por supuesto, otra clave importante. Si el nodo es de un vector, el grado representa el número de mamíferos de



(a) Los rectangulos resaltan grupos de Lutzomyia cuyas distribuciones se presentan en las otras tres imagenes

(b) Distribución geográfica de las especies marcadas con el rectángulo naranja

(c) Distribución geográfica de las especies marcadas con el rectángulo verde.

(d) Distribución geográfica de las especies marcadas con el rectángulo morado.

Figura 4.6: La topología de la red está ligada a la distribución geográfica de las características espaciales booleanas.

los que se puede estar alimentando –y por lo tanto son reservorios potenciales–, entonces un nodo vector con alta conectividad indica que es un vector que puede estar explotando muchos mamíferos y por ende puede ser clave en la propagación de la enfermedad. En el caso, si el nodo es de un mamífero y este tiene alto grado de conectividad, quiere decir que hay muchas especies de vectores que se pueden estar alimentando de éste y por lo tanto hay mayor riesgo de que sea reservorio y de que los vectores estén intercambiando el parásito a través de esta especie ([25]). Para ejemplificar, consideremos los mamíferos *Peromyscus yucatanicus*, *Ototylomys phyllotis*, *Reithrodontomys gra-*

cilis y *Heteromys gaumeri*. Estos cuatro mamíferos dieron positivo a pruebas de laboratorio para presencia del parásito de leishmaniasis, viendo sus nodos en la red –nodos resaltados en 4.4(b)–, podemos observar que son nodos conectados con la mayoría de *Lutzomyia*.

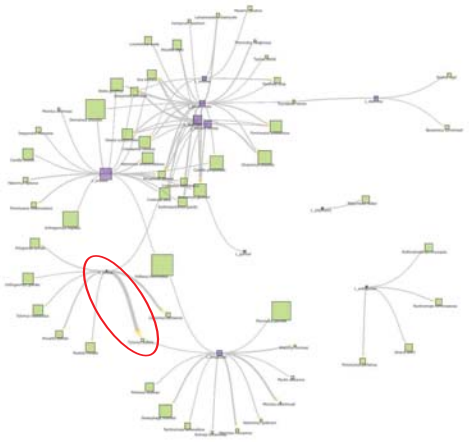


Figura 4.7: Una arista gruesa entre dos nodos pequeños muestra una relación fuerte entre dos especies poco comunes.

En resumen, este caso de estudio, muestra como las redes inferidas pueden ser herramientas útiles para el análisis exploratorio de las relaciones estadísticas entre grandes conjuntos de características espaciales booleanas. Proveen una estructura que resume propiedades del sistema que otros métodos exploratorios no logran. La estructura de la red permite comparar todos los nodos al mismo tiempo e identificar propiedades, como alto grado de conectividad, que son importantes para entender la dinámica potencial de interacciones en el sistema (ecosistema en este estudio particular). La estructura de esta red condujo a hipótesis sobre que mamíferos podrían ser reservorios –mamíferos conectados significativamente a una o más *Lutzomyia*– o que podrían cargar el parásito a otras regiones. Asimismo, vimos que la estructura de la red puede estar ligada a la distribución geográfica directamente, donde grupos de nodos densamente conectados entre sí representan regiones geográficas que son mostradas de una forma compacta ya que substraen la información espacial de la visualización y permite al analista enfocarse en la topología de las relaciones.

Aunque las redes ayudan a sintetizar la información sobre las relaciones entre un conjunto de variables, si las relaciones son muchas el análisis visual se vuelve complicado, por esta razón hay una rama de la visualización de información dedicado al análisis de redes. Uno de los temas que atraen mucho la atención es el de encontrar metodologías que ayuden al análisis mediante la construcción de agrupamientos de nodos de acuerdo a cierto parámetro de similitud. En la siguiente sección se revisará como se ha atacado esta pregunta y se propone una metodología con este fin.

4.4. Agrupamiento de nodos en redes

El análisis de redes complejas comenzó a despertar gran interés a finales del siglo XX, en parte debido a que con el crecimiento de la Red Mundial (WWW) se tuvo acceso a redes reales con una complejidad que no se había registrado antes. Esto abrió el camino para descubrir que existen propiedades que se presentan de manera recurrente en redes con orígenes completamente diferentes, ya sea en la Web, en la naturaleza, o en redes de interacciones humanas [5].

Para el estudio de redes existen varias propiedades, bien definidas, con las que se caracterizan. Por ejemplo, el diámetro, el índice de clustering o la distribución de sus conexiones (libre de escala, aleatoria), por mencionar algunas. Estas propiedades, aunque útiles, difícilmente son suficientes para entender la estructura de una red, ya que son estadísticas globales que no describen detalles sobre la estructura de la red. Para entender mejor la composición de las redes complejas, un campo que ha recibido mucha atención es la detección de comunidades.

La búsqueda de estructura comunitaria y de jerarquías es un tema de investigación activo en el campo del análisis de redes. La estructura comunitaria de una red puede dar información importante sobre su estructura, por ejemplo, en una red celular puede definir módulos con funciones específicas, o en la WWW grupos de sitios con temas similares. La estructura modular, no sólo sirve para encontrar que propiedades estructuran la conectividad de una red, también sirve como

un método para resumirla, y permite analizar como esos módulos o comunidades interactúan entre sí sin perderse en el detalle de las conexiones de cada nodo.

En el mundo se presentan redes de muchos tipos: dirigidas como una red alimenticia; no dirigidas, como una red de amigos; o dirigidas y pesadas, como una red de carreteras donde el peso representa el tiempo de recorrido. Sin embargo, la búsqueda de algoritmos para descubrir agrupamientos de nodos se ha enfocado principalmente en las redes no dirigidas y no pesadas, el caso más simple.

Además de que en la literatura se ha prestado más atención a las redes no dirigidas y sin pesos, los esfuerzos han estado centrados en el análisis de comunidades. Esta es una estructura que resulta atractiva porque sus implicaciones son quizá más fáciles de interpretar, porque explícitamente siguen principios que se presentan comúnmente en la naturaleza y en nuestra sociedad, la tendencia de las cosas a agruparse y construir sistemas más complejos a partir de grupos y de la división de tareas.

La búsqueda de agrupamientos de nodos sin sesgo hacia el tipo de estructura que se espera encontrar, permite identificar patrones de conectividad ocultos, pero que determinan la distribución de la red. Aunque en una producción menor, existen trabajos en esta dirección, un ejemplo es el algoritmo presentado en [37], el cual forma grupos de nodos de acuerdo a como se conectan con la red, por ejemplo, pueden encontrar grupos de nodos donde la propiedad que los agrupa es que favorecen las conexiones entre si (assortative mixing), pero también forma grupos donde los miembros favorecen las conexiones fuera del grupo (disassortative mixing), es decir, se conectan poco entre ellos y prefieren conectarse con nodos fuera del grupo.

En una red dirigida puede ser importante conocer grupos de nodos que están más bien determinados por quién se conecta a ellos. Por ejemplo, en una red social como Twitter, existen usuarios con una gran cantidad de seguidores, a los cuales podría ser más útil caracterizarlos según quién los sigue, al mismo tiempo habrá otros usuarios que sea mejor caracterizarlos por a quien siguen, o por ambos criterios. Sin embargo, existen pocos algoritmos que consideren tanto las conexiones de salida de

un nodo como las de entrada.

El problema de descubrir estructuras que puedan estar jugando un papel importante en la topología de una red comienza desde la definición del criterio de agrupamiento. Para definir un algoritmo de agrupamiento se requiere alguna medida que determine cuándo dos objetos son similares o distan poco el uno del otro. En el caso de las redes, si se busca agrupar nodos, o clasificarlos, hay que definir cuando dos nodos son parecidos.

Existen diversas formas de definir la similitud entre nodos, algunas de estas utilizan información como las etiquetas de estos, en otros casos sólo se considera la topología de la red, es decir, los nodos y las aristas sin atributos. En cualquier caso, la definición de similitud entre nodos está basada en alguna propiedad de conectividad, por ejemplo, número de vecinos comunes o número de caminos ajenos entre dos nodos.

En general, no se puede decir que una medida es mejor que otra, porque todo depende del problema particular. En la detección de comunidades, en general, se define que los agrupamientos están dados por grupos de nodos densamente conectados entre sí y dispersamente conectados al resto de la red. Aún así, definir cuándo un grupo de nodos están densamente conectado entre sí y dispersamente al resto de la red no es trivial y no existe un estándar. Por ejemplo, uno podría pensar que una comunidad son nodos que forman un componente completamente conectado de la red o podría dar una definición más laxa. Otro ejemplo, es que hay algoritmos que buscan una partición de la red, mientras que otros permiten que existan comunidades que se traslapan. Como se puede ver el agrupamiento de nodos en redes es un área con varios temas abiertos, que pueden ayudar a entender como se estructuran distintos tipos de sistemas a nuestro alrededor.

4.5. Medidas de similitud entre nodos

Los algoritmos de agrupamiento de nodos en en redes requieren de la definición de una medida de similitud entre nodos o aristas, o al menos del concepto que hace a dos objetos similares, esta distinción se hace porque en ocasiones no se define explícitamente una distancia entre los objetos, sino una medida que actúa sobre los agrupamientos y que indirectamente está definiendo cuando los nodos son similares.

La similitud entre dos nodos no tiene una definición única y, en general, depende de los objetivos del análisis de la red y del tipo de red. Incluso en detección de comunidades, en redes no dirigidas y sin pesos, a pesar de ser el caso más simple, existen varias medidas de similitud que se pueden utilizar y no existe un consenso sobre cual es la más adecuada. Más allá, no hay una definición estándar de comunidad.

Para el caso de redes no dirigidas y sin pesos, hay dos medidas de similitud clásicas, estas son la distancia *euclidiana* y la correlación de Pearson.

Definición 4.5.1 *La distancia euclidiana entre dos nodos n_i, n_j se define como*

$$D(n_i, n_j) = \sqrt{\sum_{k \neq i,j} (A_{ik} - A_{jk})^2} \tag{4.8}$$

donde A es la matriz de adyacencias

Cabe mencionar que el nombre de distancia euclidiana no es el más adecuado dado que el espacio de los nodos no es un espacio euclidiano.

Para definir la similitud con la correlación de Pearson primero definimos el valor esperado y la varianza de cada renglón de la matriz de adyacencia

$$\mu_i = \frac{1}{n} \sum_{j}^{n} A_{ij}, \sigma^2 = \frac{1}{n} \sum_{j}^{n} (A_{ij} - \mu_i)^2 \tag{4.9}$$

entonces tenemos la siguiente definición

Definición 4.5.2 *La correlación de Pearson entre dos nodos n_i, n_j esta dada por*

$$s(n_i, n_j) = \frac{\frac{1}{n}\sum_k^n (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{\sigma_i \sigma_j} \tag{4.10}$$

Otra medida posible para definir la similitud entre nodos sería la distancia Manhattan

Definición 4.5.3 *La distancia Manhattan entre dos nodos n_i, n_j esta dada por*

$$D(n_i, n_j) = \sum_k^n |A_{ik} - A_{jk}| \tag{4.11}$$

Como se mencionó anteriormente, existen muchas posibilidades para definir la similitud entre dos nodos. Otra opción es usar caminantes aleatorios, este método a su vez ofrece distintas formas para definir una distancia entre nodos, una de estas es definir la distancia entre dos nodos como el valor esperado del tiempo (número de pasos) que tardaría un caminante aleatorio en hacer el viaje redondo entre el nodo i y el nodo j –una de las razones por las que se toma el viaje redondo es que si sólo se considerara el viaje en un sentido no tendríamos simetría–. Esta definición tiene la ventaja de que disminuye si se aumenta el número de caminos entre los nodos o si se acorta alguno. Propiedad que no tiene, por ejemplo, la distancia por caminos más cortos [19].

Otro de los temas en detección de comunidades es cómo comparar el desempeño de los distintos algoritmos, una manera bastante estandarizada en la comunidad es usar la medida de modularidad presentada en [44], dada una partición de los nodos de una red esta medida compara la fracción de conexiones que corresponden a conexiones interiores de los grupos con el valor esperado para una red generada aleatoriamente que mantiene la valencia de los nodos en la red original.

Definición 4.5.4 *Sea R un red no dirigida y $C = \{C_1, \ldots, C_k\}$ una partición de los nodos de R*

definimos la función de modularidad como

$$M(R, C) = \sum_{i}^{k} e_{ii} - \sum_{ijk} e_i j e_k i \tag{4.12}$$

donde $e_i j$, es la fracción de aristas en la red que van del grupo C_i al grupo C_k.

En un principio la función de modularidad se presentó como una medida para calificar la estructura encontrada, pero la aplicación de esta como función objetivo en algoritmos de búsqueda es inmediata, así que esto dio pié a un nuevo tipo de algoritmos en la búsqueda de estructura comunitaria, los cuales buscan maximizar la función de modularidad, por ejemplo, usando calentamiento simulado.

En general, existe una cantidad enorme de algoritmos para el descubrimiento de agrupamientos, que utilizan técnicas muy distintas. Otro ejemplo, son los basados en inferencia bayessiana que maximizan la esperanza de que la red haya sido generada por un modelo probabilístico, dados ciertos parámetros, donde los parámetros definen cada grupo [37].

Las medidas mencionadas anteriormente, con excepción del modelo bayesiano, están definidas para redes no dirigidas, pero algunas pueden servir como base para definir similitud entre nodos de una red dirigida. Por ejemplo, existe una función de modularidad extendida para el caso dirigido [38].

Un ejemplo interesante de medida de similitud para nodos en gráficas dirigidas lo presentan en el artículo [8]. En este caso la medida de similitud cuantifica que tanto se parece la forma en que se conectan dos nodos en términos de su topología local, considerando el tipo de nodos a los que apuntan y el de los que apuntan a ellos, la medida permite comparar nodos en gráficas distintas, de tal forma que el caso en el que se comparan los nodos en la misma gráfica resulta un caso particular.

A pesar de la diversidad de medidas de similitud, la mayoría se basan en a que nodos se conecta un nodo de la red, ya sea dirigida o no, y no consideran que nodos se conectan al nodo. Este tipo de definiciones de similitud dejan de lado información potencialmente importante. Es decir, puede

haber un patrón en el tipo de nodos que se conectan a un nodo o grupo de nodos, por ejemplo si consideramos Twitter, uno probablemente puede descubrir cosas sobre un usuario por el tipo de usuarios que lo siguen y la cantidad de estos. Del mismo modo en una red alimenticia podría ser interesante que especies comparten depredadores. Esto motiva el trabajo que se presenta en la siguiente sección, en la cual se propone una medida de similitud entre nodos como un algoritmo para construir grupos de nodos que no presupone una estructura particular subyacente en la red, como es el caso cuando sólo se busca estructura comunitaria.

4.6. SOM para agrupamiento de nodos en redes

Las redes que surgen de nuestro proceso de inferencia pueden ser complejas, por lo que es necesario utilizar herramientas que ayuden al análisis. Es decir, que ayuden a descubrir patrones interesantes en la topología de la red. Un tipo de herramientas muy utilizado en el análisis de redes complejas son los algoritmos para la detección de grupos de nodos similares. El problema es que por lo general estos algoritmos se restringen a detección de comunidades –grupos de nodos densamente conectados entre sí y poco conectados al resto de los nodos en la red–. La estructura comunitaria en redes, aunque es importante, no es el único tipo de estructura que se puede presentar, ni el único que resulta interesante, en nuestro caso nos interesa un algoritmo más flexible. Es igualmente interesante saber si una red tiene una estructura bipartita, o una mezcla de comunidades y hubs, o grupos de nodos que sólo se conectan a nodos que no pertenecen al grupo. Dado que en la literatura existen pocos algoritmos de agrupamiento de nodos que tengan la flexibilidad para detectar distintos tipos de agrupamiento, y ninguno –hasta donde tenemos conocimiento– que descubra cualquier tipo de patrón de conectividad que se repita en una red, sin que este tenga que ser asociativo o disociativo, que además funcione para redes dirigidas y pesadas. Es una aportación al campo de análisis de redes desarrollar uno con estas características, que además está en línea con la filosofía

del análisis exploratorio de datos.

4.6.1. SOM para agrupamiento de nodos en redes dirigidas y pesadas

En esta sección se desarrolla una estrategia para agrupar los nodos de una red, de tal forma que estos grupos revelen el tipo de patrones de conectividad predominantes. El objetivo es que si hay, por ejemplo, estructura comunitaria predominante, los grupos resultantes deberían ser comunidades. Cuando decimos patrón de conectividad, nos referimos a un patrón representado por un conjunto de nodos tales que se conectan 'casi' a los mismos nodos y a los que se conectan también 'casi' los mismos nodos, es decir, si se toman dos nodos pertenecientes al mismo patrón la intersección de los conjuntos de vecinos debería ser cercana a la unión.[Esto se puede desarrollar en una medida de similitud por ejemplo algo como $P(E_{ik} \mid E_{jk})P(E_{jk} \mid E_{ik})$ o $P(E_{ik} \mid E_{jk})P(E_{jk} \mid E_{ik})$].

Por ejemplo, supongamos que tenemos una red bipartita no dirigida de cuatro nodos n_1, n_2, n_3, n_4, con las aristas (n_1, n_3), (n_1, n_4), (n_2, n_3), (n_2, n_4). Esta red consta de dos patrones de conectividad: $P_1 = \{n_i \mid (n_i, n_1) y (n_i, n_2)\}$ y $P_2 = \{n_i \mid (n_i, n_3) y (n_i, n_4)\}$, por lo que esperaríamos que el algoritmo encontrara los grupos $G_1 = n_1, n_2$ y $G_2 = n_3, n_4$. Lo cual en el caso del SOM equivale a que los nodos $n_1 y n_2$ sean mapeados a una neurona y los nodos $n_3 y n_4$ a otra. En este ejemplo, la red se divide perfectamente en dos grupo donde los nodos se conectan igual. Pero en la realidad los nodos que pertenezcan a un patrón se conectan parecido, más no igual. Para elaborar sobre el ejemplo anterior, agregamos un quinto nodo n_5, y la artista (n_3, n_5). ¿Qué grupo le tocaría entonces a n_5?. Si buscamos igualdad entre los miembros de un grupo, entonces n_5 requiere uno nuevo, pero si buscamos similitud entonces n_5 podría caer en G_1, ya que comparte la conexión con n_3 con los miembros de ese grupo. En cambio, con los miembros de G_2 no comparte ningún vecino. El SOM en este caso debería colocar a n_5 más cercano a n_1 que a n_3 o n_4.

Lo que buscamos con la medida de similitud es que considere tanto las conexiones de salida de

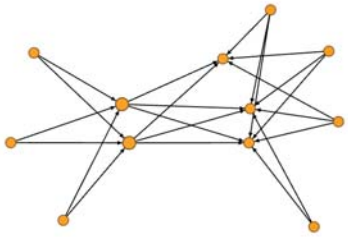


Figura 4.8: Ejemplo de una red con 4 grupos de nodos que se conectan igual

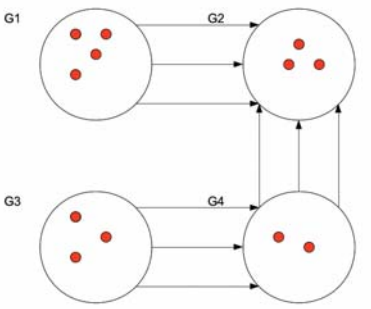


Figura 4.9: Grupos de nodos en 4.8

un nodo como las conexiones de entrada. Una forma directa es contar en cuantas conexiones de salida y en cuantas conexiones de entrada difieren. Sin embargo, otro requisito para la medida de similitud es que considere los pesos de las conexiones en el caso de redes pesadas. Por ello, una solución es usar una medida basada en la distancia euclidiana.

La propuesta combina dos cosas. Primero, una medida de similitud entre nodos que considera tanto a qué nodos se conectan como cuáles nodos se conectan a ellos. Segundo, la aplicación de un algoritmo de mapeo auto-organizado para ordenar los nodos según su similitud, lo cual corresponde a un mapeo donde nodos cercanos en la imagen son nodos que se conectan de forma muy parecida al resto de la red, esto es, se conectan más o menos a los mismos nodos y aproximadamente los mismos nodos se conectan a ellos.

El mapeo que se obtiene del SOM puede ser utilizado para visualización, por ejemplo, se puede

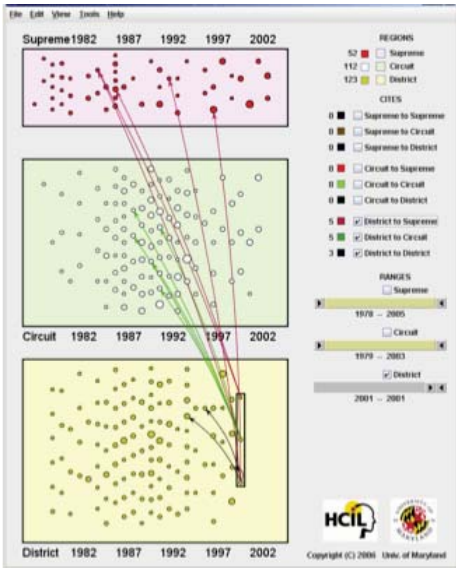


Figura 4.10: Interfaz del sistema de sustratos semánticos [4]

usar para la coloración automática de nodos, de tal manera que colores similares indiquen nodos similares o para organizar el despliegue de la red en pantalla de tal forma que nodos cercanos sean nodos parecidos; también se podría usar para el agrupamiento de nodos de tal forma que se obtenga una compactación de la red al representar agregar grupos de nodos como un sólo nodo. Por ejemplo, en [4] los autores presentan una técnica de visualización de redes, a la que llaman sustratos semánticos (ver 4.10), donde el usuario puede separar los nodos en regiones de la pantalla según alguno de sus atributos. Utilizando el SOM uno podría hacer algo similar de manera automática, según las proximidades encontradas por el SOM, como se ve en 4.13(a).

En las aplicaciones del SOM, para obtener una visualización de los datos uno puede usar mallas grandes donde podría haber más celdas que datos en la muestra, en este caso lo que importa es encontrar el ordenamiento de los datos por lo que no importa si cada uno cae en su propia celda, incluso esto es deseable en algunas ocasiones. En el caso del SOM para clustering la idea es que el mapeo encuentre grupos de datos que pertencen a una misma clase, en esta aplicación es natural que

el objetivo sea que caiga más de un elemento en las celdas del SOM. Si se hace un mapeo a un SOM con menos nodos que la red, los nodos a los que converja se pueden usar como una compactación de la red. Es decir, considerar sólo las aristas de los nodos prototipo, donde si un nodo modelo se conecta a los nodos agrupados alrededor de otro nodo modelo, entonces existe una arista entre los dos. Nótese que un nodo modelo puede ser que se conecte a varios nodos representados por otro modelo sin que esté conectado al nodo modelo en la red original, por lo tanto quizá habría que definir un umbral que determine cuando hay suficientes conexiones para que exista una arista entre dos nodos en la compactación de la red. Otra estrategia sería considerar únicamente las artistas de salida de los modelos y no las de entrada.

Además, si los nodos tienen varios atributos, los grupos que encuentre el SOM pueden ser analizados posteriormente para detectar si hay atributos que son más determinantes en la construcción de grupos. En el caso de estudio de la red ecológica del estudio de caso, las asociaciones estadísticas representadas por las aristas de la red pueden ser generadas por distintos factores, los cuales en ocasiones se desconocen y la red podría ayudar a determinarlos. Por ejemplo, si al analizar un cluster de nodos se detecta que la mayoría son especies endémicas de una zona entonces sería natural considerar que la asociación estadística que se detectó está determinada principalmente por geografía. En general, para las especies se tienen varios atributos, entre ellos sus categorías taxonómicas, pero también se pueden tener atributos como características fenotípicas. El SOM podría revelar características dominantes en la definición de como interactúan las especies.

En redes donde existe algún tipo de agente que es transportado por los nodos, el agruparlos de acuerdo a la forma en que se conectan puede permitir detectar patrones de movilidad del agente. Por ejemplo, si un grupo de nodos de animales se conecta a dos grupos que no están conectados entre sí, esto podría indicar un grupo de riesgo que sirve de puente en la transmisión de una enfermedad a otra región. Por ejemplo, en la imagen 4.9 los nodos en el grupo $G4$ sirven como puente entre los nodos en $G3$ a los nodos del $G2$.

Obviamente el significado del SOM depende del contexto de la red. Mientras que en una red ecológica, varias especies muy conectadas a un mismo conjunto de nodos, podría ser un grupo de especies depredadoras en competencia, si los nodos representan compañías que cotizan en la bolsa, este mismo patrón podría indicar compañías que dependen de un tipo de industria. Sin embargo, en el fondo el patrón podría ser el mismo, por ejemplo, competencia por un recurso.

Otra aplicación es con respecto a gráficas cuya estructura predominante es la de una gráfica k-partita, en los algoritmos de disposición de nodos por fuerzas tienden a tener problemas ya que los nodos se atraen mediante sus aristas, por lo que los nodos pertenecientes a un bloque de la gráfica se acercarán a los nodos de la red que no pertenecen a su grupo lo cual genera organizaciones de los nodos poco útiles para el usuario. Encontrar grupos de similitud y usar esos grupos para acomodar los nodos en el plano puede ser una estrategia para resolver este tipo de problemas.

En suma, este método puede ser de utilidad para detectar agrupamientos de nodos que siguen distintos patrones y que ayuden a descubrir propiedades del sistema, tipos de interacciones entre las entidades, e incluso en la toma decisiones, por ejemplo, en el control de epidemias. A continuación presentamos el algoritmo, empezando por la medida de similitud que utiliza.

4.6.2. Algoritmo

Nuestra terminología está basada en la terminología que usa Kohonen, en este párrafo la explicamos para facilitar la lectura. El SOM se compone de unidades ordenadas en una malla rectangular, donde denotaremos por u_{ij} a la unidad con coordenadas (i, j) en la malla. Cada unidad tiene asociado un elemento del conjunto de entrada –en nuestro caso nodos de la red–, el elemento asociado con una unidad es su modelo, y lo denotamos como m_{ij}, donde i, j son las coordenadas de la unidad a la que está asociado.

Dada una red, sea A su matriz de adyacencia, es decir, la matriz tal que el renglón i corresponde

a los nodos a los que se conecta el nodo n_i, y la columna i a los nodos que se conectan a éste. Definimos la distancia entre dos nodos como

$$d(n_i, n_j) = \frac{[\sum_k (A_{ik} - A_{jk})^2]^{\frac{1}{2}} + [\sum_k (A_{ki} - A_{kj})^2]^{\frac{1}{2}}}{2} \tag{4.13}$$

donde la división por 2 la hace equivalente con la definición clásica de medida euclidiana para gráficas no dirigidas.

En la sección 2.3.6 se presentó el algoritmo para el SOM no vectorial. Esta medida nos da lo que necesitamos para aplicarlo a los nodos de una red. Una ventaja del SOM no vectorial es que sólo se calculan las distancias una vez, y en el mapeo el SOM únicamente necesita sumar las distancias entre los nodos, esta es una de las características que permite hacer implementaciones eficientes.

Dada una red con n nodos, la matriz de similitud S es la matriz de $n \times n$ tal que el elemento $s_{i,j}$ corresponde a la distancia entre los nodos n_i y n_j, claramente la matriz es simétrica, con ceros en la diagonal. Esta matriz, es la entrada principal del SOM.

En las demostraciones que se presentan al final de la sección el algoritmo SOM asigna los modelos iniciales escogiendo de entre los nodos de la red aleatoriamente, y realiza la actualización utilizando una vecindad que decrece de la siguiente forma. Sea N_0 el radio inicial de la vecindad de actualización, sea t la iteración en que se encuentra el SOM y t_f el número máximo de iteraciones entonces calculamos el tamaño de la vecindad como

$$N_t = \left\lfloor N_0 - \frac{(N_0 - 1)t}{(0.75)t_f} \right\rfloor \tag{4.14}$$

Este ritmo de decrecimiento de la vecindad funcionó bien para los ejemplos que se muestran, pero es un parámetro que en general hay que determinar experimentalmente. Para actualizar un modelo m_u, necesitamos definir el peso que cada nodo de la red tiene en la determinación de la media generalizada. Sean u y u' unidades de la malla del SOM, el peso de los nodos que se mapean a u'

en la actualización de u es

$$w_u(u') = e^{-\frac{\|u-u'\|^2}{2N_t^2}} \tag{4.15}$$

Sea V el conjunto de nodos de la red, para cada nodo $n \in V$ definimos u_n como la unidad a la que es mapeado en el SOM, sea u' una unidad del SOM, la media generalizada para actualizar la unidad u' la calculamos de la siguiente manera

$$M_G(u) = argmin_{n \in V} \sum_{m \in V} w_u(u_m) D(n, m) \tag{4.16}$$

La actualización de las unidades del SOM está dada por

$$u' = M_G(u') \tag{4.17}$$

El algoritmo converge cuando los modelos al tiempo t son iguales a los modelos al tiempo $t-1$ o cuando se cumple con el número de iteraciones máximo.

4.6.3. Evaluación

La evaluación de metodologías de agrupamiento de nodos no tiene una solución estándar, por un lado ésta depende del contexto y los objetivos, por otro no se sabe que grupos existen en realidad, además, las medidas de evaluación que existen en la literatura se enfocan en agrupamientos para estructura comunitaria. El objetivo de evaluar un método de agrupamiento de nodos es saber que tan bien representan-detectan la estructura de un sistema, sin saber cuál es esa estructura –si es que existe–. Por ejemplo, en el caso de la modularidad [44], la cual se ha intentado usar como medida estándar para evaluar las comunidades que encuentra un algoritmo, se sabe que puede dar valores altos aún cuando la red es aleatoria [14].

Una técnica es evaluar con respecto a redes conocidas, donde se sabe que tienen la estructura que se espera detectar, y ver que tan bien la detecta el algoritmo. Un método parecido es utilizando redes artificiales que sean construidas con estructuras predeterminadas.

Otras técnicas se pueden tomar de la evaluación de sistemas de visualización de información. Una de ellas son los estudios con usuarios, parecido a lo que se hace en sistemas de interacción hombre-computadora, para estimar de manera cualitativa la utilidad de un sistema. Los estudios de usuarios tienen mayor costo tanto en tiempo como en infraestructura, y requieren conocimiento de técnicas en este tipo de estudios, como técnicas de evaluación de interfaces hombre-computadora que permiten evaluar cuantitativamente el desempeño del usuario al realizar ciertas tareas, pero también se requieren evaluaciones cualitativas y en contexto con el problema que se intenta resolver, para lo cual es necesaria la participación de investigadores o expertos que a través de encuestas evalúen una visualización en cuanto al tipo de descubrimientos que les permitió hacer y cómo se compara con los métodos de análisis que usan generalmente. Finalmente, otra forma de evaluar una visualización es considerando la adopción de ésta en la práctica, esta última claramente es la que más tiempo requiere y sobre la que no se tiene control alguno, por lo que resulta la menos adecuada para evaluar en el contexto de este trabajo. Por ello, nos restringimos e evaluar el algoritmo mediante métodos más controlados, para lo cual, la mejor opción es utilizar redes artificiales con estructura conocida para mostrar si el algoritmo es efectivo. Y se dejan los estudios con usuarios para trabajo futuro [Cap. 5].

En la próxima sección se presentan los resultados que obtuvimos al aplicar el algoritmo SOM para organizar los nodos de redes generadas artificialmente con estructura comunitaria, redes artificiales con estructura disociativa, y la red de mamíferos y vectores producto de nuestra metodología con épsilon.

4.6.4. Resultados del SOM para agrupamiento de nodos

Redes artificiales

El primer experimento se realizó con una red generada aleatoriamente. La red fue construida de la siguiente manera. Empezamos con 50 nodos divididos en 5 grupos disjuntos del mismo tamaño. A partir de esa división se generaron las aristas de tal modo que las artistas entre nodos del mismo grupo se generaron usando una probabilidad de 0.9 y las aristas entre nodos de distintos grupos con una probabilidad de 0.1. La estructura que se obtiene es una red pequeña con estructura comunitaria. Si calculamos la medida de modularidad de los grupos de nodos que se utilizaron para generar la red obtenemos un valor de 0.48 – recordemos que un valor de 0 indica que las conexiones entre los grupos de nodos se distribuyen como se esperaría en una distribución de aristas aleatoria y que el máximo valor de la función de modularidad es 1.0–. Este experimento nos permite comprobar que el SOM funciona en un caso básico. Dado que la red no es dirigida, la medida de similitud es igual a la distancia euclidiana entre los vectores de adyacencia. En este caso, como se puede ver en 4.11(a), la coloración de los nodos basada en el SOM identifica bien los grupos.

Para la siguiente prueba con redes aleatorias se generó una red de 100 nodos divididos en 5 grupos de 20 nodos. Los parámetros para generar esta red fueron iguales al anterior: probabilidad de 0.9 para generar aristas entre nodos del mismo grupos y probabilidad de .01 para aristas entre nodos de distintos grupos.

El SOM, definido por una malla de 10×10, convergió a la novena iteración. La vecindad de inicio fue de 4×4.

Para subir un poco el nivel de dificultad generamos una nueva red manteniendo el número de nodos pero divididos en 10 grupos, siguiendo los parámetros del experimento anterior, en este caso se obtuvo el resultado que se observa en la imagen 4.11(c). En esta imagen se puede apreciar que la coloración en algunos casos define menos claramente los grupos, esto se debe a que el número de

grupos afecta el desempeño, y entre más grupos existan, si mantenemos constante el número de nodos, resulta más difícil distinguir unos nodos de otros bajo la perspectiva de conectividad. En el caso extremo de esta metodología cada nodo pertenece a un grupo distinto, y por lo tanto, todos los nodos se conectan a otros nodos con la misma probabilidad.

La siguiente red 4.11(d) mantiene el número de nodos y las probabilidades, pero aumenta el número de grupos a 20, con 5 nodos por grupo. Como se puede ver en este caso los colores demarcan con menor claridad los grupos. Esto se nota especialmente con el color azul, el cual comparten varios nodos. Esto tiene dos factores, por un lado la desemejanza entre nodos de distintos grupos disminuye con respecto a los ejemplos anteriores, por otro, el mapa de color quizá no sea el óptimo. Por la forma en que se generó la red, existen aproximadamente 20 vectores prototipo, es decir un vector por grupo, estos además son aproximadamente equidistantes entre sí.

Redes partitas

En 4.12(a) vemos el resultado de aplicar el SOM a una red bipartita de 20 nodos, en este caso la división es perfecta. Esto no debe sorprender ya que los nodos que pertenecen al mismo grupo son idénticos de acuerdo con nuestra medida de distancia.

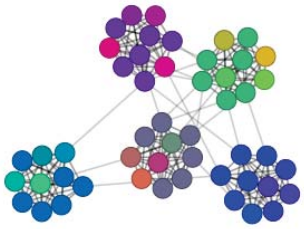
En el caso de 4.12(b) se construyó una red casi bipartita, en este caso la probabilidad de conexión entre de un nodo hacia otro nodo del mismo grupo es .1 y del nodo hacia un nodo externo es .9.

Red épsilon mamífero-vector (red pesada y dirigida)

Esta red prueba el SOM con una red pesada y dirigida. Esta es una de las principales motivaciones ya que no existen metodologías hasta donde tenemos conocimiento que consideren ambas cosas para identificar grupos de nodos.

Además demostramos otra aplicación del SOM, en la cual podemos utilizarlo para organizar la disposición de los nodos de tal forma que se organizen de acuerdo a su posición en la malla del

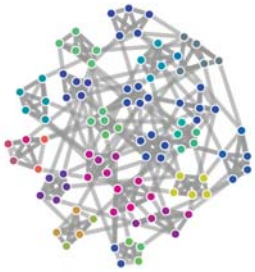
SOM. Esto nos permite identificar patrones de conectidvidad y sirve como apoyo para identificar visualmente posibles clusters de nodos, como se puede ver en 4.13(a).



(a) Coloración automática con el SOM. El algoritmo funciona bien y colorea con colores similares los miembros de cada cluster

(b) Coloración automática con el SOM para una gráfica aleatoria de 100 nodos con 5 grupos de 20 nodos. Parámetros para generar la red: Probabilidad de que un nodo tenga arista hacia otro nodo de su grupo: 0.9; Probabilidad de arista hacia un nodo de otro grupo: 0.01

(c) Coloración automática con el SOM para una gráfica aleatoria de 100 nodos con 10 grupos de 10 nodos. Parámetros para generar la red: Probabilidad de que un nodo tenga arista hacia otro nodo de su grupo: 0.9; Probabilidad de arista hacia un nodo de otro grupo: 0.01



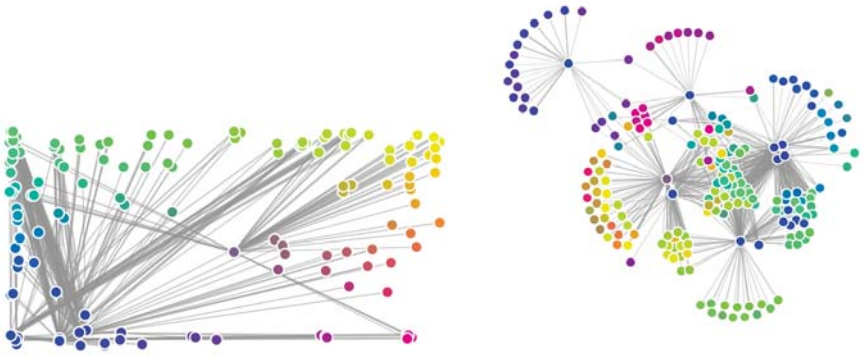
(d) Coloración automática con el SOM para una gráfica aleatoria de 100 nodos con 20 grupos de 5 nodos. Parámetros para generar la red: Probabilidad de que un nodo tenga arista hacia otro nodo de su grupo: 0.9; Probabilidad de arista hacia un nodo de otro grupo: 0.01

Figura 4.11:

(a) Coloración automática con el SOM para una gráfica bipartita de 20 nodos.

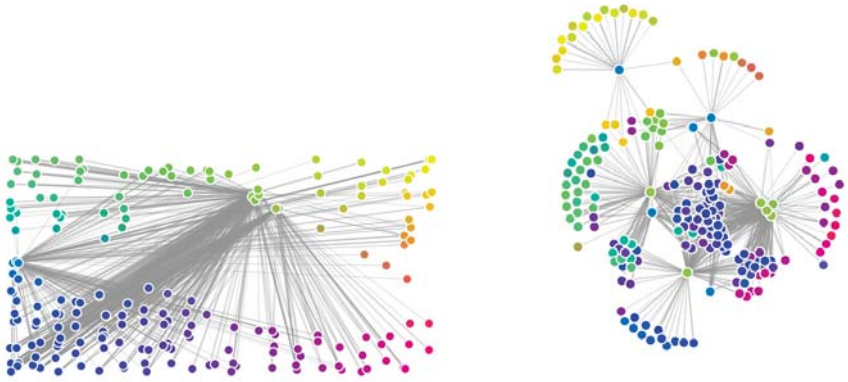
(b) Coloración automática con el SOM para una gráfica semibipartita de 20 nodos.

Figura 4.12:



(a) Coloración automática y distribución de nodos usando el SOM

(b) Coloración automática usando el SOM y distribución de nodos tradicional



(c) Coloración automática y distribución de nodos usando el SOM con normalización

(d) Coloración automática usando el SOM con normalización y distribución de nodos tradicional

Figura 4.13: Red épsilon de mamíferos con especies del género *Lutzomyia* y resultado de análisis de conectividad usando el SOM para nodos

Capítulo 5

Conclusiones y perspectivas

Esta tesis fue motivada por una idea: dada la inmensa cantidad de datos que estamos generando, necesitamos nuevos métodos para el análisis de datos que apoyen la comunicación entre distintas disciplinas y entre personas con distintos niveles de especialización. Dirigidos por dicha idea desarrollamos dos metodologías de visualización y un método de selección de nivel de granularidad para definir coincidencias entre variables espaciales. Asimismo, buscamos que este trabajo fuera más allá de la sola propuesta de metodologías y que incluyera una discusión sobre las tareas que involucra el análisis de datos en general. En dicha discusión revisamos los conceptos que motivan el análisis exploratorio y cómo se relacionan con el desarrollo de técnicas de visualización de información; elaboramos sobre lo que significa el análisis de datos y cómo el desarrollo del computo lo ha revolucionado. Nuestra opinión es que esta revolución en el análisis de datos no se limita a una cuestión cuantitativa –más datos, más rápido–. El que podamos procesar grandes cantidades de datos a gran velocidad nos ha obligado a encontrar nuevas formas de comunicar las características de los datos y de explorarlas. Asimismo, esta nuevas capacidades han motivado el desarrollo de marcos teóricos que guíen el diseño de nuevas técnicas de visualización y exploración.

Incluimos un breve recuento histórico sobre la evolución de la visualización de información co-

mo disciplina científica. La historia comienza a finales del siglo XVIII, cuando William Playfair, en un intento por resumir la gran cantidad de datos de comercio que se generaban en Gran Bretaña, inventó las gráficas de pie y de barras. Este es uno de los primeros ejemplos de visualización de información en el sentido que manejamos actualmente. A principios del siglo XX, sin embargo, el interés por los métodos de visualización de datos disminuyo considerablemente en favor de métodos cuantitativos. Esta desestimación por los métodos cualitativos, se revirtió a mediados del mismo siglo. Con la llegada de las computadoras, se renovó el interés por las técnicas de visualización, sólo que esta vez con un enfoque más formal. Jaques Bertin y John Tuckey son los primeros exponentes de este tratamiento de la visualización para la exploración de datos. Esta historia es un ejemplo de como los desarrollos tecnológicos nos motivan a su vez a construir marcos teóricos que nos permitan aprovecharlos mejor.

En el resto de la tesis propusimos soluciones a tres problemas: ¿cómo elegir el nivel de granularidad para la discretización del espacio en el análisis de datos espaciales y cuáles son las implicaciones de esta elección?; ¿cómo visualizar las correlaciones entre variables cuando el conjunto de variables es muy grande?; ¿cómo visualizar patrones de conectividad en una red de nodos y aristas?. También se presentó un caso de estudio donde se aplicó nuestra metodología: la detección de especies que podrían ser reservorios de enfermedades como la leishmaniasis.

Una característica fundamental en el análisis de datos es la simplificación de los datos. Dicha simplificación se da de varias maneras: en las características que asumimos sobre las variables, por ejemplo asumiendo independencia entre estas; en las variables que elegimos utilizar, ya sea eligiendo un subconjunto de variables o construyendo variables derivadas de las originales; y en la representación del componente de referencia, por ejemplo discretizando el espacio mediante una rejilla. Desde este punto de vista nuestra metodología ayuda a simplificar conjuntos de datos. En el capítulo 3 exploramos las implicaciones de nuestro modelos de agrupamiento de datos. En dicho capítulo analizamos como dependen los resultados de la discretización del espacio y propusimos

una técnica de optimización para guiarnos en la elección de la granularidad del modelo de agrupamiento.

En el capítulo tres llevamos a la práctica las ideas sobre la eficacia de las representaciones visuales que exponemos en los primeros capítulos. En dicho capítulo apoyamos nuestro análisis sobre los efectos de la resolución de la rejilla con dos tipos de visualización: gráficas xy tradicionales y mapas de calor. Dichas visualizaciones nos permitieron ver los cambios en la distribución de las correlaciones entre variables según la resolución de la rejilla. Dos de las aportaciones de esta tesis son herramientas visuales: visualizaciones interactivas de corrleaciones entre variables; coloración y posicionamiento automáticos de los nodos de una red, determinados por las similitudes entre las vecindades de los nodos mediante un algoritmo de aprendizaje no supervisado (Cap. 4).

En el capítulo 3 estudiamos el problema de la unidad de área modificable, un problema del que se tiene registro desde principios del siglo XX, y que fue largamente ignorado porque se creía irresoluble. En los 80 este problema fue retomado por algunos investigadores en ciencias geográficas. En esta tesis nos enfocamos al problema cuando la división del espacio esta definida por una partición regular, como una rejilla. Nos acercarnos al problema de tres formas: mediante simulaciones; desde la probabilidad; y analizando datos reales. El problema, como vimos, se complica cuando estamos analizando muchas variables porque para cada pareja puede haber una resolución óptima particular. Además de la discusión del problema, en este capítulo proponemos elegir la rejilla de manera automática maximizando el promedio de ε para todos los pares de variables.

En el último capítulo 4, utilizamos las técnicas de análisis de datos espaciales del capítulo anterior para construir una visualización que permita al analista representar y estudiar el sistema de relaciones estadísticas que existen en un conjunto de datos espaciales multivariado. Esta metodología busca la simplificación de un sistema a través de un método de inferencia y su presentación visual en forma de red, la cual es una representación natural e intuitiva de sistemas de relaciones que permite al analista tener una imagen completa de las interacciones en el sistema, y le permite explorar

estas relaciones a varios niveles de significatividad estadística de manera interactiva.

Al sustraer la información de la distribución de las variables sobre los componentes de referencia, nuestra herramienta de visualización le da al analista una perspectiva que le permite concentrarse en la topología de las relaciones. Asimismo, aunque se sustrae la representación explícita de la distribución espacial la topología de la red sigue ligada a esta. En el caso de estudio mostramos un ejemplo de que cuando la relación entre multiples variables está ligada a una zona del espacio de referencia entonces esta zona se reflejara en la red como un cúmulo de nodos.

La metodología integra naturalmente tres tipos de filtros: selección de tipos de nodos, tipo de relaciones y pesos de las aristas. Dichos filtros permiten la exploración de los datos desde distintos puntos de vista y distintos niveles de complejidad. La metodología además es modular. Por ejemplo, podemos cambiar el modelo de co-ocurrencia o la medida estadística y el proceso sería el mismo en esencia. También podemos integrar nuevas variables de referencia como altitud y/o tiempo. Por ejemplo, si tenemos altitud además de latitud y longitud, podríamos considerar celdas tridimensionales para el modelo de co-ocurrencia. Nótese que no tienen que ser cubos, se puede definir una granularidad deiferente para cada eje. Análogamente, podríamos incluir una variable temporal. Dimensiones extra, por supuesto, hacen el análisis más complejo, y la elección de las dimensiones de las celdas más complicado.

En la segunda parte del capítulo 4 presentamos una técnica para colorear los nodos de una red de tal forma que nodos con vecindades similares tengan asignados colores similares. Dicha coloración no es trivial porque los vectores de adyacencia de los nodos de la red son de alta dimensión. Para definir la coloración, proyectamos los vectores de adyacencia de los nodos a las celdas de una retícula de dos dimensiones mediante un mapeo auto-organizado las celdas de la retícula a su vez tienen colores asignados de tal forma que celdas cercanas tienen colores parecidos. Finalmente, a cada nodo se le asigna el color de la celda a la que es proyectado su vector de adyacencia. Dicha técnica nos ayuda a identificar patrones de conectividad entre los nodos de la red.

5.1. Trabajo futuro

En el desarrollo de esta tesis encontramos algunas direcciones para mejorar nuestro sistema de visualización. Quedó pendiente diseñar una interfaz de visualización más completa. Aunque una parte de esto involucra implementar métodos de interacción estándar. Hay otra parte que corresponde al diseño de vistas ligadas para explorar la conexión entre la distribución espacial de las variables y la topología de la red. Una de las tareas sería filtrar desde una perspectiva lo que se observa en otra perspectiva. Específicamente nos referimos a dos perspectivas: la perspectiva espacial (en el caso de estudio el espacio geográfico) y la de redes. Por ejemplo, el usuario podría restringir los datos a una región espacial y que la red se actualice correspondientemente. O seleccionar nodos en la red y que se muestren únicamente las distribuciones espaciales de las variables correspondientes. En esta área existen trabajos sobre interfaces con vistas ligadas que pueden ser tomados como punto de partida [12].

Otra forma de utilizar la red sería para explorar la sensibilidad de la estadística de correlación a la resolución de la rejilla. El analista ganaría intuición al explorar interactivamente los cambios en la estructura de la red de correlaciones según se hace más fina o más gruesa la rejilla que define coincidencias entre variables.

Consideramos que el método de optimización de rejilla se puede afinar. La idea es utilizar un algoritmo que combine los tres elementos que utilizamos en el capítulo 3: optimizar la ecuación que calcula el número esperado de celdas con coincidencias entre las variables, asumiendo que estas siguen una distribución aleatoria uniforme; refinar la resolución que determinó la fase previa maximizando el número de celdas con coincidencia; finalmente, refinar aún más la resolución que encuentra la segunda fase maximizando ε. La razón para la primer fase es que es más eficiente optimizar la ecuación del valor esperado que optimizar el número de coincidencias para cada resolución. Dicha ecuación, sin embargo, asume una distribución aleatoria uniforme. Por lo tanto

utilizamos el método más costoso de calcular efectivamente el número de coincidencias para refinar la optimización, por último, como vimos, optimizar el número de coincidencias puede tener problemas cuando los conjuntos de puntos tienen tamaños muy desiguales, por eso se utilizaría como fase final la optimización de ε directamente.

El ordenamiento de los nodos por el SOM se podría usar de otras formas además de las que presentamos al final del capítulo 4. En dicho capítulo, los nodos se trasladaron a las posiciones de sus unidades correspondientes en el SOM. Creemos que se podría adaptar el algoritmo de dibujo de redes dirigido por fuerzas (*force-directed graph drawing) tradicional para que las distancias entre las unidades del SOM sirvan como una fuerza adicional de atracción entre nodos. También ubicamos un problema en la coloración del SOM. En la versión actual aún cuando el SOM es pequeño se confunden los colores de celdas distintas. Se requiere definir una mejor función de color que represente con mejor definición las diferencias entre unidades. Finalmente, otra forma de utilizar el SOM sería agrupar los nodos que caen en una misma unidad en un sólo nodo, con lo cual observaríamos como se conectan los nodos prototipo entre sí. El primer problema que habría que resolver para dicho mecanismo es definir cuándo dos nodos prototipo se unen por una arista, ya que puede suceder que no sean consistentes las adyacencias entre nodos prototipo.*

Bibliografía

[1] Gennady Andrienko, Natalia Andrienko, Urska Demsar, Doris Dransch, Jason Dykes, Sara I. Fabrikant, Mikael Jern, Menno J. Kraak, Heidrun Schumann, and Christian Tominski. Space, time and visual analytics. Int. J. Geogr. Inf. Sci.*, 24(10):1577–1600, October 2010.*

[2] Natalia Andrienko and Gennady Andrienko. Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer, 1 edition, December 2005.*

[3] L. Antiqueira, M. Nunes, O. Oliveirajr, and L. Fcosta. Strong correlations between text quality and complex networks features. Physica A: Statistical and Theoretical Physics*, 373:811–820, January 2007.*

[4] Aleks Aris and Ben Shneiderman. Network visualization by semantic substrates. IEEE Transactions on Visualization and Computer Graphics*, 12(5), 2006.*

[5] Albert-Laszlo Barabasi. Linked: How Everything Is Connected to Everything Else and What It Means*. Plume, April 2003.*

[6] Jacques Bertin. Semiology of Graphics: Diagrams, Networks, Maps*. ESRI Press, 1 edition, November 2010.*

[7] Julian Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.*

[8] Vincent Blondel, Anahi Gajardo, Maureen Heymans, Pierre Senellart, and Paul Van Dooren. A measure of similarity between graph vertices. Jul 2004.

[9] Andreas Buja, Deborah F. Swayne, Michael L. Littman, Nathaniel Dean, Heike Hofmann, and Lisha Chen. Data visualization with multidimensional scaling. Journal of Computational and Graphical Statistics*, 17(2):444–472, 2008.*

[10] Ramon F. Cancho and Richard V. Solé. The small world of human language. Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265, November 2001.*

[11] Deepayan Chakrabarti and Christos Faloutsos. Graph mining: Laws, generators, and algorithms. ACM Comput. Surv.*, 38(1):2, 2006.*

[12] Christopher Collins and Sheelagh Carpendale. VisLink: revealing relationships amongst visualizations. IEEE Transactions on Visualization and Computer Graphics*, 13(6):1192–9, 2007.*

[13] Kevin M. Curtin. Network analysis in geographic information science: Review, assessment, and projections. Cartography and Geographic Information Science*, 34(2):103–111, April 2007.*

[14] Leon Danon, Albert D'iaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.*

[15] Urška Demšar, A. Stewart Fotheringham, and Martin Charlton. Exploring the spatio-temporal dynamics of geographical processes with geographically weighted regression and geovisual analytics. Information Visualization*, 7:181–197, June 2008.*

[16] Ugur Dogrusoz, Konstantinos G. Kakoulis, Brendan Madden, and Ioannis G. Tollis. On labeling in graph visualization. Information Sciences*, 177(12):2459–2472, 2007.*

[17] Philip Edmonds. Choosing the word most typical in context using a lexical co-occurrence network. In Proceedings of the 35th Annual Meeting of the Association for Computational

Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 507–509, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.*

[18] Arthur Flexer. On the use of self-organizing maps for clustering and visualization. Intell. Data Anal.*, 5(5):373–384, 2001.*

[19] Francois Fouss, Alain Pirotte, Jean-michel Renders, and Marco Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, March 2007.*

[20] Jerome H. Friedman and Werner Stuetzle. John w. tukey's work on interactive graphics. The Annals of Statistics*, 30(6):1629–1639, 2002.*

[21] Michael Friendly. Milestones in the history of thematic cartography, statistical graphics, and data visualization. August 2009.

[22] M. T. Gastner and Newman. The spatial structure of networks. The European Physical Journal B - Condensed Matter and Complex Systems*, 49(2):247–252, January 2006.*

[23] C. E. Gehlke and Katherine Biehl. Certain effects of grouping upon the size of the correlation coefficient in census tract material. Journal of the American Statistical Association*, 29(185), 1934.*

[24] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Proceedings of the National Academy of Sciences*, 99:7821–7826, 2002.*

[25] Camila González, Constantino González-Salazar, Joaqu'in Heau, Carlos Ibarra-Cerdeña,

Victor Sánchez-Cordero, and Christopher R. Stephens. Using Biotic Interaction Networks for Prediction in Biodiversity and Emerging Diseases. PLoS ONE*, 4(5), May 2009.*

[26] D. Guo. Visual analytics of spatial interaction patterns for pandemic decision support. International Journal of Geographical Information Science*, 21(8):859, 2007.*

[27] Diansheng Guo. Flow mapping and multivariate visualization of large spatial interaction data. IEEE Transactions on Visualization and Computer Graphics*, 15(6):1041–1048, 2009.*

[28] Berthold Michael/ Hand. Intelligent Data Analysis*. Springer, 2nd edition, February 2007.*

[29] D. J. Hand, Heikki Mannila, and Padhraic Smyth. Principles of data mining*. MIT Press, 2001.*

[30] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. The elements of statistical learning: data mining, inference, and prediction*. Springer, 2001.*

[31] Nathalie Henry, Jean-Daniel Fekete, and Michael J. McGuffin. NodeTrix: a hybrid visualization of social networks. IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, 2007.*

[32] Weidong Huang, Peter Eades, and Seok-Hee Hong. Measuring effectiveness of graph visualizations: A cognitive load perspective. Information Visualization*, 8(3):139–152, 2009.*

[33] Dieter Jungnickel. Graphs, Networks and Algorithms (Algorithms and Computation in Mathematics)*. Algorithms and computation in mathematics. Springer, 1999.*

[34] Michael Kaufmann, Marc Kreveld, and Bettina Speckmann. Subdivision drawings of hypergraphs. In Graph Drawing: 16th International Symposium, GD 2008, Heraklion, Crete, Greece, September 21-24, 2008. Revised Papers*, pages 396–407. Springer-Verlag, 2009.*

[35] Teuvo Kohonen and Timo Honkela. Kohonen network. Scholarpedia*, 2(1):1568, 2007.*

[36] E. A. Leicht and M. E. J. Newman. Mixture models and exploratory analysis in networks. Proc Natl Acad Sci U S A*, 2007.*

[37] E. A. Leicht and M. E. J. Newman. Mixture models and exploratory analysis in networks. Proc Natl Acad Sci U S A*, May 2007.*

[38] E. A. Leicht and M. E. J. Newman. Community structure in directed networks. Physical Review Letters*, 100(11):118703, March 2008.*

[39] Luciano, Francisco A. Rodrigues, Gonzalo Travieso, and Villas P. R. Boas. Characterization of complex networks: A survey of measurements. http://arxiv.org/abs/cond-mat/0505185, 2005.

[40] Omid Madani and Jiye Yu. Discovery of numerous specific topics via term co-occurrence analysis. In Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1841–1844, New York, NY, USA, 2010. ACM.*

[41] Camilla Magnusson and Hannu Vanharanta. Visualizing sequences of texts using collocational networks. In Petra Perner and Azriel Rosenfeld, editors, Machine Learning and Data Mining in Pattern Recognition*, volume 2734 of* Lecture Notes in Computer Science*, chapter 24, pages 291–304. Springer Berlin / Heidelberg, Berlin, Heidelberg, June 2003.*

[42] Elise Desmier Nazha Selmaoui, Frédéric Flouvat and Dominique Gay. A clustering-based visualization of spatial patterns. Technical report, University of New Caledonia, 2009.

[43] M. E. J. Newman. Fast algorithm for detecting community structure in networks. Phys. Rev. E*, 69(6):066133, 2004.*

[44] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Phys. Rev. E*, 69(026113), 2003.*

[45] S. Openshaw. The modifiable areal unit problem. *Concepts and techniques in modern geography. Geo Books, 1983.*

[46] Adam Perer and Ben Shneiderman. Balancing systematic and flexible exploration of social networks. IEEE Transactions on Visualization and Computer Graphics*, 12(5):693–700, October 2006.*

[47] Alasdair Rae. From spatial interaction data to spatial interaction information? geovisualisation and spatial structures of migration from the 2001 UK census. Computers, Environment and Urban Systems*, 33(3):161–178, May 2009.*

[48] Shashi Shekhar and Yan Huang. Discovering Spatial Co-location Patterns: A Summary of Results. Lecture Notes In Computer Science*, 2121, 2001.*

[49] Jonathon Shlens. A tutorial on principal component analysis. In Systems Neurobiology Laboratory, Salk Institute for Biological Studies*, 2005.*

[50] Ben Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. ACM Trans. Graph.*, 11(1):92–99, January 1992.*

[51] R. Sierra and C. R. Stephens. Exploratory analysis of the interrelations between co-located boolean spatial features using network graphs. International Journal of Geographical Information Science*, 26(3):441–468, February 2012.*

[52] Devinderjit Sivia and John Skilling. Data Analysis: A Bayesian Tutorial*. Oxford University Press, USA, 2 edition, July 2006.*

[53] J Thomas and K Cook. Illuminating the Path: The RD Agenda for Visual Analytics*. IEEE Computer Society, 2005.*

[54] Edward R. Tufte. Beautiful Evidence. *Graphics Pr, 1st edition edition, 2006.*

[55] John W. Tukey. Exploratory Data Analysis. *Addison Wesley, 1 edition, January 1977.*

[56] Anne Veling and Peter Van Der Weerd. Conceptual grouping in word co-occurrence networks. In Proceedings of the 16th international joint conference on Artificial intelligence - Volume 2*, pages 694–699, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.*

[57] Juha Vesanto. SOM-Based data visualization methods. INTELLIGENT DATA ANALYSIS*, 3, 1999.*

[58] Joe Whittaker. Graphical Models in Applied Multivariate Statistics. *Wiley Publishing, 2009.*

[59] Wikipedia. Florence nightingale — wikipedia, the free encyclopedia, 2012.

[60] Feng Xie and David Levinson. Measuring the structure of road networks. Geographical Analysis*, 39(3):336–356, July 2007.*

[61] Hui Xiong. Encyclopedia of GIS. *Springer, February 2008.*

[62] Hujun Yin. ViSOM - a novel method for multivariate data projection and structure visualization. IEEE Transactions on Neural Networks*, 13(1):237–243, January 2002.*

[63] Hujun Yin. Nonlinear Multidimensional Data Projection and Visualisation. In Intelligent Data Engineering and Automated Learning*, volume 2690 of* Lecture Notes in Computer Science*, pages 377–388. Springer Berlin / Heidelberg, 2003.*

[64] Hujun Yin. On multidimensional scaling and the embedding of self-organising maps. Neural Networks*, 21(2-3):160–169, March 2008.*

[65] Hujun Yin. The Self-Organizing maps: Background, theories, extensions and applications. In John Fulcher and L. Jain, editors, Computational Intelligence: A Compendium*, volume 115 of* Studies in Computational Intelligence*, pages 715–762. Springer Berlin / Heidelberg, 2008.*

[66] Jin S. Yoo and Shashi Shekhar. A joinless approach for mining spatial colocation patterns. IEEE Transactions on Knowledge and Data Engineering*, 18(10):1323–1337, 2006.*

[67] Ilmi Yoon, Sanghyuk Yoon, Neo Martinez, Rich Williams, and Jennifer Dunne. Interactive 3D visualization of highly connected ecological networks on the WWW. In Proceedings of the 2005 ACM symposium on Applied computing*, pages 1207–1212, Santa Fe, New Mexico, 2005. ACM.*

[68] Xin Zhang, Nikos Mamoulis, David W. Cheung, and Yutao Shou. Fast mining of spatial collocations. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 384–393, Seattle, WA, USA, 2004. ACM.*