UNIVERSIDAD DE GRANADA

Máster Universitario en Estadística Aplicada

Trabajo de fin máster

ESTIMACIÓN DEL GASTO POR DIABETES DE UN PROSPECTO EN UNA PÓLIZA DE GASTOS MÉDICOS COLECTIVO, MEDIANTE MODELOS DE REGRESIÓN ESPACIAL

Autor: Cindy Teresa Mazariegos Ochoa

Tutor: José Luis Romero Béjar

Febrero 2023

Capítulo 1

Introducción

1.1 Introducción

La enfermedad de Diabetes es una enfermedad crónica degenerativa. Cuando una persona tiene la glucemia elevada, se la conoce como hiperglucemia y se deriva de una diabetes no controlada, que con el tiempo ocasiona daños graves a varios órganos y sistemas del cuerpo, sobre todo los nervios y los vasos sanguíneos. Debido a estas complicaciones, la Diabetes es una enfermedad significativa para una aseguradora. El sistema estadístico del sector asegurador (SESA) informa que, dentro de los padecimientos más frecuentes del año 2020, se encuentran, en el número 10, las enfermedades endocrinas, nutricionales y metabólicas, siendo la Diabetes una enfermedad de este tipo. El siniestro más alto por monto acumulado es el de *Otros trastornos endocrinos, nutricionales y metabólicos* que asciende $115,830,000.00 y que seguía abierto para el año 2020. Un *Siniestro de Diabetes* con duración de 29 años se encuentra en el quinto lugar de los siniestros con mayor duración históricamente, con un gasto promedio pagado por año de $39,383 en el estado de Jalisco.

Estos resultados destacan como el costo de siniestralidad en los seguros de Gastos Médicos Mayores y de Salud van en aumento durante el pasar de los años, y en particular, para el padecimiento de Diabetes nos habla de que es de las enfermedades mas frecuentes, costosas y duraderas.

Actualmente, las aseguradoras predicen los gastos de un prospecto con la enfermedad de Diabetes por medio de consultas a un Doctor especialista que a través de su experiencia y conocimiento determina un importe estimado, sin embargo, esta técnica no se sustenta con modelos estadísticos y hoy en día no se tienen estudios al respecto que propongan un modelo para esta predicción deseada.

En este trabajo se pretende buscar un modelo de regresión espacial que permita estimar el gasto que generará un prospecto de un seguro de Gastos Médicos Mayores Colectivo, permitiendo a la aseguradora determinar el costo del seguro que sea rentable y competitivo.

Este trabajo de fin de máster, en primer lugar, presentará información sobre qué es la Diabetes, información sobre las cifras de Diabetes en la República Mexicana, en qué estados los encuestados admitieron ser diagnosticados de Diabetes y en qué porcentaje son hombres o mujeres. Posteriormente este trabajo profundizará sobre los Seguros de Gastos Médicos mayores y cómo han impactado las reclamaciones por Diabetes a lo largo del tiempo. En la segunda parte, se presentarán los aspectos metodológicos y fundamentales para el análisis espacial, como calcular la correlación espacial y los modelos de regresión espacial. En la tercera parte, se analizarán los datos de una aseguradora mexicana, que contienen los importes pagados por siniestros de diabetes del año 2008 al 2021 y, aprovechando que se cuenta con una variable espacial que es el estado de la república mexicana donde ocurre, se realizará un análisis exploratorio de datos, se analizará la correlación espacial de los estados de los importes pagados medios y la duración media, para así determinar si es posible encontrar un modelo de regresión espacial significativo. Por último, se presentarán los resultados y conclusiones obtenidos en el análisis.

1.2 Objetivo

El objetivo fundamental de este trabajo es realizar un análisis estadístico de los gastos por siniestros del padecimiento de Diabetes que tiene una aseguradora mediante modelos de regresión espacial. Este análisis puede ser útil para plantear una alternativa a la estimación del gasto por diabetes que tendrá un asegurado de una póliza de gastos médicos colectivo en un año concreto.

En este sentido, para la suscripción de Gastos Médicos Mayores Colectivo, es importante una adecuada estimación de complementos de un padecimiento, para poder otorgar una prima de seguro rentable y competitiva, por lo que se pretende realizar un análisis exploratorio de datos, incluyendo la variable espacial y encontrar un modelo de regresión espacial que nos permita estimar el complemento del siniestro de Diabetes para suscribir un negocio nuevo.

Capítulo 2

Marco Teórico

2.1 Diabetes

La diabetes sacarina o diabetes mellitus es una enfermedad crónica. Esta enfermedad se ocasiona cuando el páncreas no produce la insulina suficiente para el cuerpo o bien el organismo no la utiliza eficazmente. La insulina es una hormona encargada de regular la cantidad de glucosa en la sangre, a esta medida de concentración de glucosa en la sangre se le llama Glucemia. Cuando una persona tiene la glucemia elevada, se le conoce como hiperglucemia y se deriva de una diabetes no controlada, que con el tiempo ocasiona daños graves a varios órganos y sistemas del cuerpo, sobre todo los nervios y los vasos sanguíneos (Organización Mundial de la Salud, 2021). Para simplificar, en el presente trabajo nos referiremos a la diabetes mellitus como “diabetes”.

Existen tres tipos principales de diabetes: diabetes tipo 1 que es la diabetes insulinodependiente, juvenil o de inicio en la infancia, diabetes tipo 2 no insulinodependiente o de inicio en la edad adulta y diabetes gestacional.

La diabetes de tipo 1 es cuando hay una producción deficiente de insulina y se administra diariamente insulina al afectado. Hasta la fecha de este tipo de diabetes no se conoce su causa o como prevenirla y para el 2017 se tenían 9 millones de personas con esta enfermedad, la mayoría de estos son de paises de renta alta. (Organización Mundial de la Salud, 2021).

La diabetes de tipo 2 es cuando el organismo no utiliza eficazmente la insulina. Mayormente este tipo de diabetes es la que presentan las personas, mas de un 95 %, se atribuye altamente al exceso de peso y la falta de actividad fisica. Hoy en día este tipo de diabetes no solo se presenta en adultos, ahora también cada vez más en niños. (Organización Mundial de la Salud, 2021).

La diabetes gestacional se da durante el embarazo, los niveles de glucosa son altos pero menores al parametro para diagnosticar diabetes. Este tipo de diabetes aumenta el riesgo de complicaciones en el embarazo y el parto. Además, la madre y sus hijos serán propensos a adquirir diabetes tipo 2 en el futuro. (Organización Mundial de la Salud, 2021).

A nivel mundial la FID[1] estima que en 2019 había 463 millones de personas con diabetes y que esta cifra puede aumentar a 578 millones para 2030 y a 700 millones en 2045[2]. De acuerdo con el comunicado de prensa 645/21 del 12 de noviembre de 2021 hecho por el INEGI[3], en México, durante 2018 de acuerdo con la Encuesta Nacional de Salud y Nutrición había 82,767,605 personas de 20 años o más en el país, de las cuales 10.32% reportaron contar con un diagnóstico médico previo de diabetes mellitus. Por sexo, 13.22% de las mujeres de 20 años o más disponían de este diagnóstico y 7.75% en los hombres de 20 años o más. Es decir, la enfermedad está más presente en las mujeres que en los hombres (INEGI, 2021).

Este mismo comunicado (Centro Estadístico del Sector Asegurador, 2020) nos menciona que se detectó que a medida que aumenta la edad de las personas se presenta un incremento de la enfermedad de diabetes; a nivel nacional el 25.8% de las personas con edad 60 a 69 años informaron tener diabetes (2.3 millones). En este rango de edad las mujeres representaron un 35.6 % con diabetes (1.4 millones). A partir de 70 años se presenta más la enfermedad en hombres con un 18.4% (714 mil) (INEGI, 2021).

En el siguiente mapa de la republica méxicana (ver Figura 1) se visualiza por estado el porcentaje de prevalencia de diabetes, donde los estados de Coahuila, Nuevo Leon, Tamaulipas, Campeche, Hidalgo y Ciudad de México se destacan por tener mayor porcentaje de población de 20 años o más, que al encuestar declararon tener diagnostico de diabetes.

[1] Federación Internacional de Diabetes (2019). Versión Online del Atlas de la Diabetes de la FID. Novena edición 2019. pág. 4 https://www.diabetesatlas.org/upload/resources/material/20200302_133352_2406-IDF-ATLAS-SPAN-BOOK.pdf

[2] Federación Internacional de Diabetes (2019). Versión Online del Atlas de la Diabetes de la FID. Novena edición 2019. pág. 4 https://www.diabetesatlas.org/upload/resources/material/20200302_133352_2406-IDF-ATLAS-SPAN-BOOK.pdf

[3] INEGI: Instituto Nacional de Estadística y Geografía de México

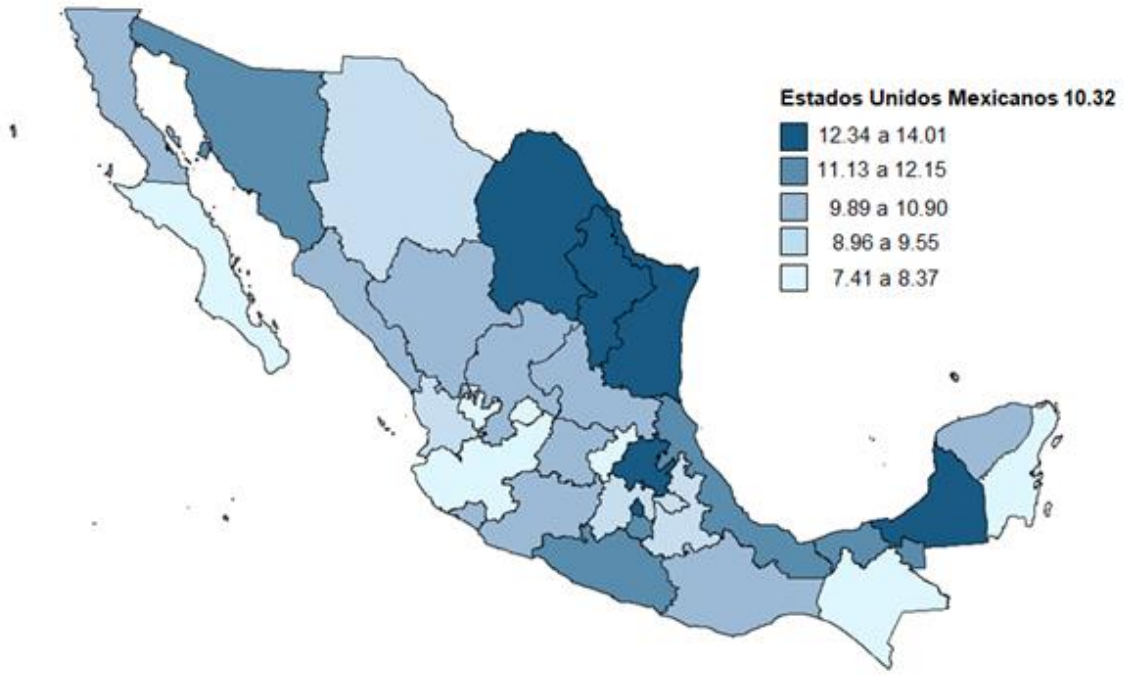


Figura 1: Porcentaje de la población de 20 años o más que al momento de la encuesta declaró tener un diagnóstico previo de diabetes. Fuente: INEGI, INSP, SALUD. Encuesta Nacional de Salud y Nutrición 2018.

2.2 Seguros de Gastos Médicos Mayores

Es importante mencionar algunos conceptos acerca de seguros, de acuerdo con AMIS[4], el seguro de Gastos Médicos Mayores es un contrato que tiene como objetivo la protección para hacer frente a los gastos médicos originados por un accidente o enfermedad (AMIS, s.f.). En el mercado asegurador se tienen los seguros de Gastos Médicos Mayores Colectivos dirigidos a las empresas y organizaciones legalmente establecidas, mediante los cuales el empleador podrá proteger a sus empleados, ya que cuenta con una cobertura que les resarza de las pérdidas económicas no previstas, derivadas de accidentes y enfermedades (AMIS, s.f.).

El asegurado es la persona amparada en el contrato de seguro y un siniestro se define como la ocurrencia de una enfermedad o accidente cubierta en el seguro de acuerdo con las condiciones generales del contrato (AMIS, s.f.).

La suscripción de un seguro colectivo se refiere al análisis del riesgo para la cobertura de los empleados de una empresa, donde se establecen las condiciones del contrato del seguro y el costo; el costo es la prima de la póliza del seguro.

[4] AMIS: Asociación Mexicana de Instituciones de Seguros

Cuando una aseguradora suscribe ciertos negocios, cuenta con información histórica propia de siniestralidad, esta información es de años pasados, por ejemplo, de un año, 2 o 5 años de experiencia de siniestralidad. Esta información permite conocer las diferentes enfermedades que se han reclamado, cuando y el importe pagado a la fecha.

Para los seguros de Gastos Médicos Mayores Colectivo, las aseguradas generalmente recurren a un médico especialista para conocer si por las enfermedades pasadas ya reclamadas se seguirán generando gastos en el próximo año y una estimación del importe faltante en dicho gasto, a esto se le conoce como complemento de un siniestro, pero esta estimación que realizan los médicos es en base a su experiencia y conocimiento acerca de los costos de tratamientos médicos y quirúrgicos, no mediante un modelo matemático. De poder establecer un modelo que realice esta predicción, una aseguradora podría ahorrar tiempo operacional y encontrar predicciones más cercanas a la realidad.

El sistema estadístico del sector asegurador (SESA), es el centro estadístico donde AMIS concentra y brinda información estadística sobre los seguros. En su reporte de Accidentes y Enfermedades del 2020 podemos encontrar resultados interesantes (SESA, 2020), como:

- 96% de los gastos generados en los seguros se concentra en el ramo de Accidentes y Enfermedades y Salud (ver Figura 2).

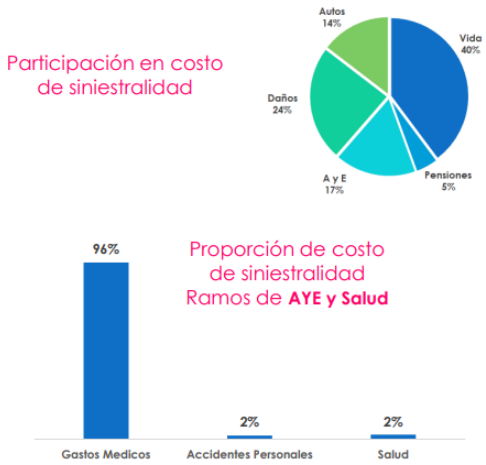


Figura 2: Porcentaje de gastos por ramo de seguros, SESA 2020.

- Para el año 2020 el costo de siniestros en Gastos Médicos Mayores fue de $66,700,000,000.00 (ver Figura 3).

Figura 3: Costo de siniestros Gastos Médicos Mayores del 2003 al 2020

- Dentro de los padecimientos más frecuentes del año 2020, se encuentran, en el número 10, las enfermedades endocrinas, nutricionales y metabólicas, siendo la Diabetes una enfermedad de este tipo (ver Figura 4).

Centro Estadístico del Sector Asegurador

SEGURO DE GASTOS MÉDICOS
PADECIMIENTOS MAS FRECUENTES 2020

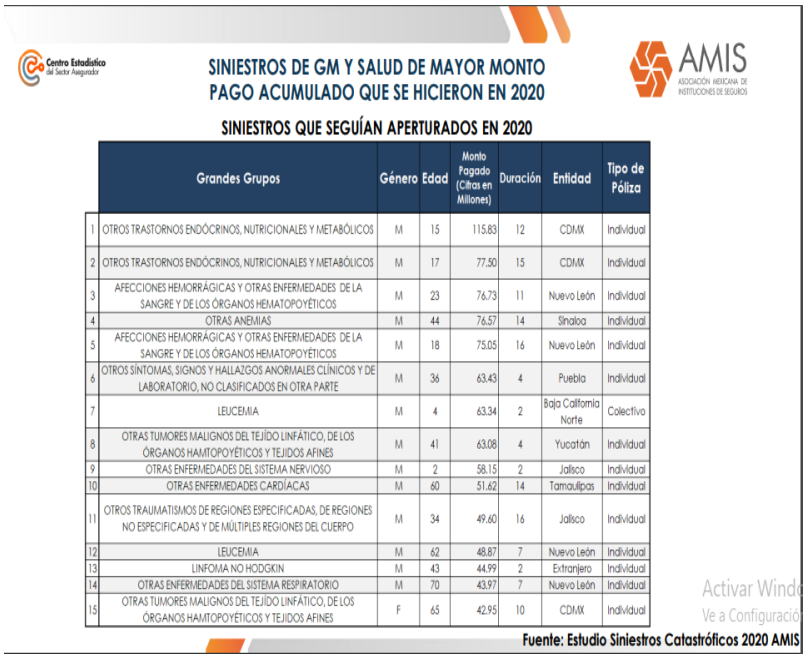
AMIS Asociación Mexicana de Instituciones de Seguros

Montos Mayores a $5,000 pesos

	Capítulos CIE-10	Monto Siniestros
1	NEOPLASIAS	8,103
2	ENFERMEDADES DEL SISTEMA OSTEOMUSCULAR Y DEL TEJIDO CONECTIVO	4,348
3	ENFERMEDADES DEL APARATO DIGESTIVO	3,722
4	ENFERMEDADES DEL SISTEMA CIRCULATORIO	3,346
5	TRAUMATISMOS, ENVENENAMIENTOS Y ALGUNAS OTRAS CONSECUENCIAS DE CAUSA EXTERNA	3,322
6	ENFERMEDADES DEL APARATO GENITOURINARIO	2,122
7	ENFERMEDADES DEL SISTEMA RESPIRATORIO	1,896
8	ENFERMEDADES DEL SISTEMA NERVIOSO	1,329
9	EMBARAZO, PARTO Y PUERPERIO	1,133
10	ENFERMEDADES ENDOCRINAS, NUTRICIONALES Y METABÓLICAS	1,009
11	CIERTAS ENFERMEDADES INFECCIOSAS Y PARASITARIAS	796
12	CIERTAS AFECCIONES ORIGINADAS EN EL PERIODO PERINATAL	778

Figura 4: Padecimientos mas frecuentes 2020.

- El siniestro más alto por monto acumulado es el de Otros trastornos endocrinos, nutricionales y metabólicos (ver Figura 5) por $115,830,000.00 y que seguía abierto para el año 2020. Por regla general este tipo de siniestro se refiere a una persona de sexo masculino, con edad de 15 años al iniciar el padecimiento, con duración de 12 años, del estado de Ciudad de México.



SINIESTROS DE GM Y SALUD DE MAYOR MONTO PAGO ACUMULADO QUE SE HICIERON EN 2020

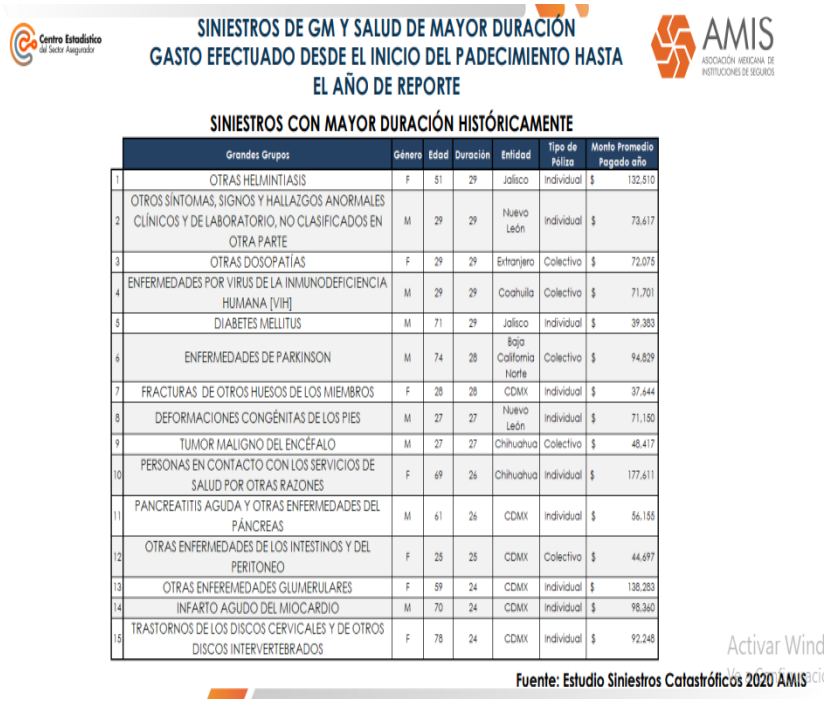
SINIESTROS QUE SEGUÍAN APERTURADOS EN 2020

	Grandes Grupos	Género	Edad	Monto Pagado (Cifras en Millones)	Duración	Entidad	Tipo de Póliza
1	OTROS TRASTORNOS ENDÓCRINOS, NUTRICIONALES Y METABÓLICOS	M	15	115.83	12	CDMX	Individual
2	OTROS TRASTORNOS ENDÓCRINOS, NUTRICIONALES Y METABÓLICOS	M	17	77.50	15	CDMX	Individual
3	AFECCIONES HEMORRÁGICAS Y OTRAS ENFERMEDADES DE LA SANGRE Y DE LOS ÓRGANOS HEMATOPOYÉTICOS	M	23	76.73	11	Nuevo León	Individual
4	OTRAS ANEMIAS	M	44	76.57	14	Sinaloa	Individual
5	AFECCIONES HEMORRÁGICAS Y OTRAS ENFERMEDADES DE LA SANGRE Y DE LOS ÓRGANOS HEMATOPOYÉTICOS	M	18	75.05	16	Nuevo León	Individual
6	OTROS SÍNTOMAS, SIGNOS Y HALLAZGOS ANORMALES CLÍNICOS Y DE LABORATORIO, NO CLASIFICADOS EN OTRA PARTE	M	36	63.43	4	Puebla	Individual
7	LEUCEMIA	M	4	63.34	2	Baja California Norte	Colectivo
8	OTRAS TUMORES MALIGNOS DEL TEJIDO LINFÁTICO, DE LOS ÓRGANOS HAMTOPOYÉTICOS Y TEJIDOS AFINES	M	41	63.08	4	Yucatán	Individual
9	OTRAS ENFERMEDADES DEL SISTEMA NERVIOSO	M	2	58.15	2	Jalisco	Individual
10	OTRAS ENFERMEDADES CARDÍACAS	M	60	51.62	14	Tamaulipas	Individual
11	OTROS TRAUMATISMOS DE REGIONES ESPECIFICADAS, DE REGIONES NO ESPECIFICADAS Y DE MÚLTIPLES REGIONES DEL CUERPO	M	34	49.60	16	Jalisco	Individual
12	LEUCEMIA	M	62	48.87	7	Nuevo León	Individual
13	LINFOMA NO HODGKIN	M	43	44.99	2	Extranjero	Individual
14	OTRAS ENFERMEDADES DEL SISTEMA RESPIRATORIO	M	70	43.97	7	Nuevo León	Individual
15	OTRAS TUMORES MALIGNOS DEL TEJIDO LINFÁTICO, DE LOS ÓRGANOS HAMTOPOYÉTICOS Y TEJIDOS AFINES	F	65	42.95	10	CDMX	Individual

Fuente: Estudio Siniestros Catastróficos 2020 AMIS

Figura 5: Siniestros de GM y Salud de mayor monto acumulado al 2020.

- Un Siniestro de Diabetes con duración de 29 años se encuentra en el quinto lugar de los siniestros con mayor duración históricamente (ver Figura 6), con un gasto promedio pagado por años de $39,383 del estado de Jalisco.



SINIESTROS DE GM Y SALUD DE MAYOR DURACIÓN GASTO EFECTUADO DESDE EL INICIO DEL PADECIMIENTO HASTA EL AÑO DE REPORTE

SINIESTROS CON MAYOR DURACIÓN HISTÓRICAMENTE

	Grandes Grupos	Género	Edad	Duración	Entidad	Tipo de Póliza	Monto Promedio Pagado año
1	OTRAS HELMINTIASIS	F	51	29	Jalisco	Individual	$ 132,510
2	OTROS SÍNTOMAS, SIGNOS Y HALLAZGOS ANORMALES CLÍNICOS Y DE LABORATORIO, NO CLASIFICADOS EN OTRA PARTE	M	29	29	Nuevo León	Individual	$ 73,617
3	OTRAS DOSOPATÍAS	F	29	29	Extranjero	Colectivo	$ 72,075
4	ENFERMEDADES POR VIRUS DE LA INMUNODEFICIENCIA HUMANA [VIH]	M	29	29	Coahuila	Colectivo	$ 71,701
5	DIABETES MELLITUS	M	71	29	Jalisco	Individual	$ 39,383
6	ENFERMEDADES DE PARKINSON	M	74	28	Baja California Norte	Colectivo	$ 94,829
7	FRACTURAS DE OTROS HUESOS DE LOS MIEMBROS	F	28	28	CDMX	Individual	$ 37,644
8	DEFORMACIONES CONGÉNITAS DE LOS PIES	M	27	27	Nuevo León	Individual	$ 71,150
9	TUMOR MALIGNO DEL ENCÉFALO	M	27	27	Chihuahua	Colectivo	$ 48,417
10	PERSONAS EN CONTACTO CON LOS SERVICIOS DE SALUD POR OTRAS RAZONES	F	69	26	Chihuahua	Individual	$ 177,611
11	PANCREATITIS AGUDA Y OTRAS ENFERMEDADES DEL PÁNCREAS	M	61	26	CDMX	Individual	$ 56,155
12	OTRAS ENFERMEDADES DE LOS INTESTINOS Y DEL PERITONEO	F	25	25	CDMX	Colectivo	$ 44,697
13	OTRAS ENFEREMEDADES GLUMERULARES	F	59	24	CDMX	Individual	$ 138,283
14	INFARTO AGUDO DEL MIOCARDIO	M	70	24	CDMX	Individual	$ 98,360
15	TRASTORNOS DE LOS DISCOS CERVICALES Y DE OTROS DISCOS INTERVERTEBRADOS	F	78	24	CDMX	Individual	$ 92,248

Fuente: Estudio Siniestros Catastróficos 2020 AMIS

Figura 6: Siniestros de GM y Salud con mayor duración historicamente.

Estos resultados nos hablan de como el costo de siniestralidad en los seguros de Gastos Médicos Mayores y de salud van en aumento durante el

pasar de los años, y en particular para el padecimiento de Diabetes nos habla de que es de las enfermedades mas frecuentes, costosas y duraderas. Es por esto que el interés de este trabajo se concentra en buscar un modelo que permita predecir el gasto que se generará a futuro, para una persona que es parte de un colectivo con el padecimiento de Diabetes, usando la información que generalmente se cuenta al suscribir un negocio como: cantidad pagada a la fecha, fecha de inicio de padecimiento, fechas de pago, sexo, edad y estado en donde vive, más otras variables que pudieran surgir al momento de realizar el análisis.

Capítulo 3

Métodos Estadísticos

3.1 Introducción a la Econometría y sus Aplicaciones

El significado etimológico de Econometría es medición de la economía, eco del griego oiko-nomos, que es economía, y metría, medición.

La econometría busca modelar, explicar y medir las relaciones cuantitativas y/o cualitativas de variables económicas, con ayuda del conocimiento de teorías económicas, matemáticas y estadísticas, es decir, en la Econometría se aplican métodos estadísticos y matemáticos a datos económicos, con la finalidad de comprender por qué se producen e intentar pronosticar lo que puede pasar en el futuro, por ejemplo, proyectar ventas, predecir los precios, demanda, tasa de inflación, etc. (David E. Rodriguez Guevara, 2017)

Los modelos que se aplican en la Econometría requieren de conocer teorías económicas, para facilitar la comprensión y relacionar variables, y así encontrar causales de un problema de interés (Luis Quintana Romero, 2017).

Aris Spanos menciona que “La econometría se interesa por el estudio sistematico de fenomenos económicos utilizando datos observables” (Spanos, 1996, p.3).

De esta definición se resalta la diferencia mas significativa que existe entre otros campos de la economía ya que en la econometría se utilizan datos observables, se analizan esos datos y se investiga cómo estan relacionados, para poder resolver problemas de interes económico.

Despues de analizar los datos y conocer como se relacionan las variables, se busca la función que mejor represente esta relación, dando lugar al modelo de regresión lineal, que involucraría a la variable dependiente (y) y variables explicativas (x_i) de la forma (Luis Quintana Romero, 2017):

$y_i = \beta_0 + \beta_1 x_{i1} + \dots \beta_k x_{ik} + \varepsilon$ con i en un conjunto finito de índices.

Donde, $\beta_0, \dots, \beta_k$ son parámetros desconocidos a estimar por el modelo y ε es el error.

La base de la econometría esta dada con el modelo clásico de Gauss. En este se establecen los siguientes supuestos: linealidad, independencia, homocedasticidad, normalidad y no colinealidad, el número de observaciones n debe ser mayor que el número de parametros por estimar y la naturaleza de las variables X (varianza positiva y sin valores atípicos) (Damodar N. Gujarati, 2009).

Estos supuestos nos ayudan a aplicar el modelo al problema de interés que deseemos estudiar, pero, ¿qué sucede si dentro de nuestras variables tenemos información sobre la ubicación? Al tener una relación espacial nos enfrentamos a problemas de efectos espaciales, como son la dependencia espacial y la heterogeneidad espacial. Esta última nos dice que la relación varía a medida que nos movemos en el espacio, por lo que no se cumpliría el supuesto del modelo de Gauss, que indica que debe existir una relación lineal con varianza constante. Esto nos lleva a la necesidad de buscar procedimientos alternativos para modelar, surgiendo así una especialidad en econometría que es econometría espacial.

En el presente trabajo se pretende estimar el complemento del gasto por padecimiento de Diabetes de una persona asegurada en un seguro de gastos médicos colectivos y para ello contamos con ciertas caracteristicas como: la edad, sexo, fecha de ocurrido, fecha de pagado, monto pagado y lugar de residencia. Al existir una varible de ubicación se esperaría encontrar una influencia de la regionalidad. Es decir, que el gasto por Diabetes se relaciona con los usos, construmbres, provocando posiblemente mayor o menor complicación, lo que implicaría un gasto mayo o menor.

3.2 Econometría Espacial

La econometría espacial se considera que nace en el año 1979 cuando los profesores Jean Paelinck y Leo Klaassen publican Spatial Econometrics, mencionando en su trabajo la necesidad de sistematizar la rama econométrica para los modelos regionales y urbanos (Paelinck & Klaassen 1979, p.166). No obstante, no fueron los primeros en advertir esta necesidad, para los años cincuenta, el desarrollo del Análisis Regional de Walter Isard también hacía ver la necesidad de una rama de econometría espacial (Aroca, 2000).

Por lo que, la econometría espacial nace de esta necesidad de diferenciarse de la econometría tradicional.

Teniendo en cuenta entonces que, la econometría espacial es una rama de la econometría general que, además de aplicar estadística a datos económicos, incluye también el tratamiento de datos espaciales para estimar parámetros económicos.

La econometría espacial utiliza técnicas de contraste, estimación y predicción para el análisis de datos espaciales.

La econometría espacial brinda enfoques alternativos de estimación debido a que nos encontramos con dos problemas que surgen cuando los datos incluyen ubicación: la dependencia y la heterogeneidad espaciales. La dependencia porque las variables explicativas no son fijas, estas dependerán de su ubicación por la influencia de vecinos y, la heterogeneidad porque la relación varía a medida que nos movemos en el espacio.

De acuerdo con (ANSELIN, 2001), la econometría espacial es aquella que estudia la autocorrelación espacial y la heterogeneidad espacial en los modelos de regresión de corte transversal y de datos panel.

3.2.1 Datos Espaciales

Primero conoceremos qué son los datos espaciales y sus tipos, esto nos ayudará a entender el desarrollo estadístico para los tipos de datos a los que nos enfrentamos.

Los datos espaciales son aquellos que tienen asociada una ubicación geográfica (por ejemplo, coordenadas o código postal), y que podemos integrar con una variable de tiempo y una variable descriptiva de la entidad (por ejemplo, temperatura o el producto interno bruto de la región, etc.).

(Cressie, 1993) nos define los datos espaciales como un proceso estocástico espacial. Este es un conjunto de variables aleatorias $\mathbf{y_i}$, en la ubicación **i,** donde i se encuentra en un subconjunto del espacio euclidiano D ($D \subset \mathbb{R}^2$) : $\{y_i, i \in D\}$.

Los datos espaciales pueden ser Geostadísticos, Aerales o regionales y de Patrón de puntos, esto dependerá del supuesto del subconjunto del espacio euclidiano D que hayamos dado. Los datos Geostadísticos son datos que se originan de una superfice continua dentro de un subconjunto del espacio euclidiano D, como por ejemplo si se estuviera interesado en la cantidad de precipicatión que cae en cada localización espacial de toda la superficie de México, donde D es el area de México. Los datos regionales son datos que se originan de un índice D fijo, donde un conjunto de puntos o áreas regulares o irregulares dividen el espacio euclidiano D, por ejemplo cuando se divide el área de México con datos del producto interno bruto por estados de la república Méxicana, típicamente en econometría espacial se analizan los datos regionales. Ahora los datos de patrón de puntos son puntos en el espacio euclidiano D originados por un proceso aleatorio, ya que su finalidad es localizar un caso en particular (Hildegart, 2018).

En el párrafo anterior dimos la clasificación de los datos espaciales dependiendo del supuesto dado para D, a continuación veremos tres tipos de objetos espaciales con la que se expresa la ubicación **i** (la localización)

- Los **puntos:** la localización se expresa mediante coordenadas terrestres de latitud y longitud, esto es, un punto donde se puede ubicar una empresa, un individuo o una ciudad**.** (Yrigoyen, 2003).
- Las **líneas:** en este objeto espacial la localización está dada por una línea con cierta distancia o más bien arcos porque la tierra es esférica, un ejemplo serían las calles de una ciudad. (Yrigoyen, 2003)
- Los **polígonos:** son aquellas figuras planas que forman un área, se forman conectando líneas o puntos, por ejemplo, los estados de la república mexicana. (Yrigoyen, 2003)

3.2.2 Visualización de Datos Espaciales

Uno de los pasos previos al modelado de datos es realizar un análisis exploratorio de los datos espaciales, por lo que representar estos datos en un mapa nos ayudará a analizar el área.

El mapa coroplético es el más utilizado, éste es un mapa que usa colores o patrones para identificar cierta escala sobre el valor que se tenga de la variable de interés. El criterio para establecer las clases e intervalos varía de acuerdo con el investigador, pero generalmente recomiendan la siguiente formula: 1+ 3.3ln(n), donde n es el número de áreas (Sánchez, 2018). En cuanto a la selección del intervalo se tienen los siguientes esquemas (Figura 7):

- **Divisiones naturales:** Cuando la división se puede asignar deductivamente de acuerdo con limites que se sabe son relevantes o con ayuda de la herramienta del sistema GIS[5]. (Díaz, 2018)
- **Divisiones por cuantiles:** En esta división las clases se generan con un número igual de observaciones, por ejemplo: Cuartiles (cuatro categorías), quintiles (cinco categorías). (Díaz, 2018)
- **Divisiones de intervalos iguales:** En este caso las observaciones se distribuyen uniformemente en su rango, pero, en este caso, nos enfrentamos a tener un gran número de observaciones en unas pocas clases si los datos son altamente sesgados.
- **Divisiones según desviación estándar:** Los intervalos se distribuyen alrededor de la media en unidades de desviación estándar. (Díaz, 2018)

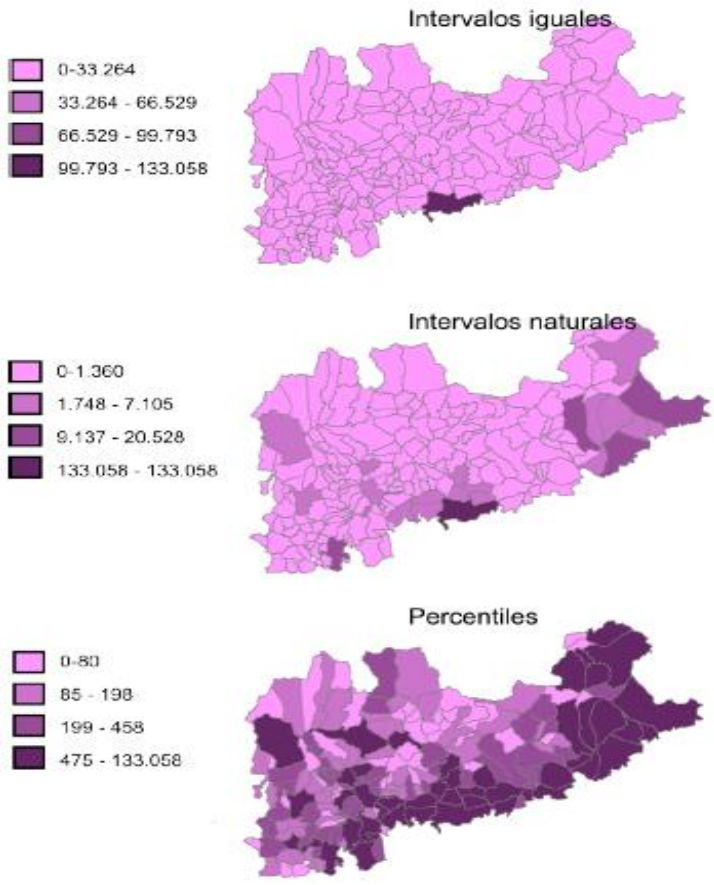


Figura 7: Tipos de intervalos de clases (Díaz, 2018).

[5] GIS, son las siglas del Sistema de Información Geográfica. Es un conjunto de herramientas que permiten la organización, almacenamiento, análisis y modelización de grandes cantidades de datos procedentes del mundo real que están vinculados a una referencia espacial.

3.2.3 Matriz de Pesos Espaciales

Como hemos mencionado, al analizar datos que cuentan con una variable espacial nos podemos enfrentar a una dependencia espacial, esta dependencia nos hablaría de una relación para localizaciones vecinas o bien una relación para localizaciones que se encuentren más lejanas, esto es dependencia espacial positiva o negativa, respectivamente. Para poder medir la relación entre las variables y su localización se utiliza la matriz de pesos espaciales, ya que brinda una medida de similitud de localizaciones, esta matriz la denominaremos **W**, donde cada elemento w_{ij} son los pesos espaciales que representan la vecindad; cuando w_{ij} son uno significa que interactúan entre sí, en caso contrario sería cero (Luis Quintana Romero, 2017).

Al tener nuestro dato espacial referenciado en un mapa, podemos identificar sus fronteras por lo que nos ayuda a establecer sus vecinos, por ejemplo, las localizaciones que se expresan en polígonos ya tienen establecidas sus fronteras y sus vecinos serian aquellos polígonos que comparten frontera. La matriz de vecindad por contigüidad se construye por sus fronteras en común, las opciones para las medidas principales de vecindad son tipo torre, tipo alfil y tipo reina, como los nombres de las piezas de ajedrez porque es semejante a su movimiento en el tablero. (Luis Quintana Romero, 2017).

En el tipo torre, la vecindad entre los puntos está dada al norte, sur, este y oeste. En el tipo alfil, su vecindad está dada por su diagonal, que sería noreste, sureste, noroeste y suroeste. Y en el tipo reina, su vecindad es hacia todos los lados, norte, sur, este, oeste, noreste, sureste, noroeste y suroeste (Hildegart, 2018)

Para la matriz de vecindad por distancia, además de considerar la vecindad física, se considera la interacción entre regiones distantes, por lo que en esta matriz W, la medida de vecindad se construye con algún criterio de distancias, como vecindad por distancia más corta, por ejemplo, cuando tenemos datos de área. La distancia entre los polígonos se mide con los centroides de los polígonos o el criterio de los k vecinos más cercanos, donde se elige de cada vecino el punto más cercano hasta obtener el número de vecinos establecidos k (Hildegart, 2018)

3.2.4 Autocorrelación Espacial

Comenzaremos citando algunas definiciones de autocorrelación espacial que han dado diferentes autores.

Para Cliff y Ord (1973), la autocorrelación espacial busca ver si la distribución de cierta variable en los estados de un país influye con más o menos probabilidad en los estados vecinos. En tal caso, significaría que existe autocorrelación espacial. Sokal y Oden (1978) mencionan que el análisis de

autocorrelación espacial determina si el valor observado de una variable es independiente a los valores en los estados vecinos. Upton y Fingleton (1985) lo explican como el atributo que tienen los datos cuando se identifica una similitud en su comportamiento.
Estos autores mencionan que la autocorrelación espacial se da cuando hay una variación espacial sistemática en los valores a lo largo del área estudiada.
Goodchild (1987) y Griffth (1991) resaltaron que la autocorrelación espacial maneja al mismo tiempo la información de ubicación y la de atributos. Goodchild (1987) menciona que la autocorrelación espacial es el nivel en que los valores localizados en un estado son similares a otros estados cercanos, aludiendo a la ley de geografía que dice “todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes” (Tobler, 1970).

Dadas las referencias anteriores de autocorrelación espacial, concluimos que si el valor de una o varias variables en una ubicación son similares a los valores de dichas variables en ubicaciones cercanas, entonces se dirá que se tiene autocorrelación espacial positiva, por el contrario, se dirá que se tiene autocorrelación espacial negativa, cuando más cerca en el espacio implique que los valores son más diferentes y menos diferentes cuando más lejos.

La autocorrelación cero será cuando los valores de las variables no dependen de su ubicación, en este caso los datos no serían espaciales, por lo que no serían abordados en este trabajo.

Como hemos visto para construir la matriz de pesos se puede utilizar el criterio de contigüidad, ésta se necesita para el análisis de autocorrelación espacial. La medida de contigüidad es la relación de vecindad, en la Figura 8 se ilustran los tipos de contigüidad. Son nombrados de la forma indicada porque se asemejan a los movimientos que hacen las figuras de ajedrez:

- Caso de torre: la contigüidad es por vecinos de 4 ubicaciones (Norte, Sur, Este y Oeste) adyacente a cada celda.
- Caso de alfil: la contigüidad es solo en las diagonales de cada celda (Noreste, Sureste, Noroeste y Suroeste).
- Caso de reina: la contigüidad se considera un vecindario de ocho celdas (Norte, Sur, Este, Oeste, Noreste, Sureste, Noroeste y Suroeste.

Figura 8: Tipos de relaciones de vecindad.

Por tanto, la relación de vecindad determina como los individuos se relacionan entre sí. Por ejemplo, si usamos pesos binarios, quedaría con $w_{ij}=1$ cuando i y j cumplen con el criterio de vecindad elegido y $w_{ij}=0$ cuando no. En el caso del criterio de distancia, los pesos binarios pueden reemplazarse por alguna función de distancia entre unidades espaciales. Un ejemplo es, usar la función inversa $w_{ij}=f(d_{ij})= 1 / d_{ij}$, tal que a mayor distancia menor relación entre i y j, cumpliendo esta forma con la ley de Tobler. Otro ejemplo de función de distancia podría definir cada peso como $w_{ij} = [|z_i-z_j|+1]^{-1} *(1/d_{ij})$, donde z es una variable que captura una distancia socio-económica (PBI, flujo comercial, etc.), en otros casos se aplica directamente $w_{ij} = [|z_i-z_j|+1]^{-1}$, donde z representa el ingreso per cápita por ejemplo (Hildegart, 2018).

Los vecinos contiguos de primer orden se definen como áreas que tienen un límite común:

$$W_{ij} = \begin{cases} 1 \text{ si el área j comparte un límite común con el área i} \\ 0 \text{ caso contrario} \end{cases}$$

Otra alternativa es que dos áreas i y j pueden unirse como vecinas cuando la distancia d_{ij} entre sus centroides es menor que un valor crítico dado, pongamos d, donde las distancias se calculan a partir de la información sobre latitud y longitud, s(i); de las ubicaciones del centroide:

$$Wij = \begin{cases} 1 \text{ si } d_{ij} < d, (d<0) \\ 0 \text{ caso contrario} \end{cases}$$

A continuación, en la Tabla 1 se presenta un resumen con algunas especificaciones para la matriz de pesos espaciales sugeridas por diferentes autores:

Referencia	Modelo	Descripción
Dacey (1968)	$w_{ij} = d_{ij} \cdot \alpha_i \cdot \beta_{i(j)}$	. d_{ij}: distancia entre los puntos o regiones (i, j) . α_i: proporción de i sobre el área total de regiones. . $\beta_{i(j)}$: proporción del perímetro de i en contacto con j
Cliff y Ord (1973)	$w_{ij} = d_{ij}^{-a}\left[\beta_{i(j)}\right]^b$	. a, b: parámetros positivos
Bodson y Peeters (1975)	$w_{ij} = \sum_{n=1}^{N} K_n \left\{ \frac{a}{1 + b \cdot e^{-c_j d_{ij}}} \right\}$	. K_n: importancia del medio de comunicación n . N: total de medios de comunicación considerados . a, b, c_j: parámetros a estimar.
Anselin (1980)	$w_{ijt} = d_{ij}^{-2}$	
Cliff y Ord (1981)	$w_{ij} = \left(c + d_{ij}\right)^{-a}$	. c: término constante positivo
Case *et al.* (1993)	$w_{ij} = \frac{1}{\left\lvert x_i - x_j \right\rvert}$	. x: variable socioeconómica (ej., PIB per cápita).
Molho (1995)	$w_{ij} = \frac{E_j^{-ad_{ij}}}{\sum_{k \neq i} E_k^{-ad_{ik}}} \;;\; \forall i \neq j$	. E: volumen de empleo
Ma *et al.* (1997)	$w_{ij} = e^{-d_{ij}^a} \quad w_{ij} = (l_{ij}/l_i)^a \quad w_{ij} = \frac{(l_{ij}/l_i)^a}{d_{ij}^{-b}}$	. l_{ij}: longitud de frontera entre las regiones (i,j) . l_i: perímetro de la región i
Toral (2000A,B)	$w_{ij} = \delta_{ij} \frac{k_i k_j p_i p_j}{d_{ij}^a}$	. δ_{ij} = 1, si las unidades espaciales i, j tienen una frontera en común y cero, si no la tienen. . p: población; k: longitud (km) de carreteras . d_{ij}: distancia por carretera entre las capitales de i, j . a: parámetro positivo, con valores 0, 1 ó 2.
Van der Kruk (2001)	$W = \sum_{d=1}^{D} W_d$	. d: orden de vecindad . D: número máximo de órdenes de vecindad existentes

Tabla 1: Modelos para matriz de pesos. Fuente: (Yrigoyen, 2003).

3.3 Medidas de Autocorrelación Espacial

La medida de autocorrelación espacial se puede utilizar para sondear la existencia de relación espacial entre distintas características de la población. Existen dos tipos de medidas: medidas globales y medidas locales.

Las medidas globales se plasman en un indicador de autocorrelación espacial o similitud general de regiones. Para resumir la zona de un estudio utilizamos un estadístico del análisis espacial global donde se asume homogeneidad, permitiendo que resulte un valor para cualquier matriz de pesos espaciales.

Las medidas locales, se determinan para cada región, permiten resolver si la región está rodeada de regiones con valores altos o bajos de estas medidas o si es similar o diferente a las regiones vecinas. Los valores de las estadísticas

locales no se generalizan para toda el área analizada y se trabaja con información sobre la posición de cada región en relación con los vecinos.

3.3.1 Medidas Globales

Las estadísticas globales pueden identificar conglomerados y relaciones espaciales solo para todo el sistema. Sin embargo, estas estadísticas se pueden desagregar en estadísticas locales, ayudándonos a detectar patrones de relaciones espaciales locales entre las regiones y sus vecinos.

Cuando estudiamos la autocorrelación espacial con una perspectiva global, el objetivo es contrastar tendencias o la distribución de una variable en el espacio o bien contrastar la hipótesis de que existen relaciones similares o distintas con regiones vecinas (Yrigoyen, 2003).

Las dos medidas más utilizadas para la autocorrelación espacial son el estadístico I de Moran y el estadístico C de Geary (Moran, 1950). La medida I de Moran viene dada por la siguiente expresión:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

Donde n es el número de áreas que consideramos, Wij la matriz de pesos, z_i y z_j el valor de la variable Z del área i y j respectivamente y $\bar{z}$ la media aritmética de todos los valores de las áreas estudiadas. (Moran, 1950)

La interpretación está dada por lo siguiente:

- Si I>0: autocorrelación espacial positiva, los valores de observación a la distancia *d* son similares.
- Si I<0: autocorrelación espacial negativa, los valores de observación a la distancia *d* son diferentes.
- Si I=0: los valores de observación a la distancia *d* se distribuyen aleatoriamente.

El índice C de Geary, propuesto por Geary en 1954, es un índice de comparaciones por pares de datos entre las diferentes áreas y viene dada por la siguiente formula:

$$c = \frac{(n-1)}{2\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(z_i - z_j)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

Donde n, Wij, z_i, z_j y $\bar{z}$ son los valores descritos igual que en la medida de I de Moran.

La interpretación está dada por lo siguiente:

- 0<C<1: autocorrelación espacial positiva.
- 1<C<2: autocorrelación espacial negativa.

3.3.2 Medidas Locales

El fenómeno de dependencia local es cuando en un sitio del espacio global que estamos analizando se identifica una concentración de valores, ya sean altos o bajos, en comparación con el valor medio de la misma, esto es conocido como “puntos calientes (hotspots)”, “puntos fríos”, “picos”, “bolsas” o “valores atípicos” de una variable (Yrigoyen, 2003).

Anselin (1995) propuso indicadores locales de relaciones espaciales (*local indicators of spatial association* - LISA). LISA[6] incluye estadísticos locales: I_i de Moran y C_i de Geary.

La diferencia entre el estadístico de Moran local y el Geary es que Moran permite la identificación de efectos de aglomeración espacial, mientras que el Geary local muestra similitudes y diferencias espaciales.

Las estadísticas locales de Moran brindan información sobre grupos de valor bajo o alto, mientras que los estadísticos locales de Geary muestran diferencias promedio entre el objeto y los vecinos, lo que ayuda a encontrar valores atípicos y patrones de similitud/diferencia (Morales-Oñate, 2022)

LISA, ayuda a identificar los puntos calientes, es decir, centros con valores altos rodeados de valores bajos, así como clústeres locales en ausencia de autocorrelación global. Las islas de alto valor pueden interpretarse no solo como puntos calientes, sino también como valores atípicos. Entonces, los estadísticos locales son un indicador de inestabilidad y desviaciones locales del patrón de autocorrelación global (Morales-Oñate, 2022).

En resumen, los LISA nos presentan dos componentes:

- Un estadístico para cada unidad espacial que puede ser sujeto a pruebas de significancia.
- Cuantifica la relación entre el estadístico global y el local: la suma de los estadísticos locales es proporcional al estadístico global.

[6] LISA: (Local Indicator of Spatial Association), son estadísticas que descomponen el índice global de autocorrelación y verifica en cuanto contribuye cada unidad espacial a la formación del valor general, permitiendo capturar de forma simultánea el grado de asociación espacial y la heterogeneidad resultante del aporte de cada unidad espacial.

El estadístico local: I Moran viene dado por:

$$I_i = \frac{(x_i - \bar{x}) \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - x)^2/n}$$

Donde los símbolos usados en la definición son similares a los estadísticos antes descritos.

La interpretación está dada por lo siguiente:

- Si I>0: autocorrelación espacial positiva, los valores de observación a la distancia *d* son similares.
- Si I<0: autocorrelación espacial negativa, los valores de observación a la distancia *d* son diferentes.
- Si I=0: los valores de observación a la distancia *d* se distribuyen aleatoriamente.

Para evaluar la significancia se evalúan los valores:

- p−value<α: El resultado será significativo.
- p−value>α: El resultado no será significativo.

3.4 Modelos de Regresión Espacial

La econometría espacial se caracteriza principalmente por ver como los efectos espaciales son tomados en cuenta. Con ayuda del análisis exploratorio antes visto podremos plantear los modelos que relacionan las observaciones de una variable con las observaciones de otras variables que se encuentran en cada región o estado del área analizada.

Para los modelos de regresión espacial se utilizan datos obtenidos de un diseño transversal simple. Se toma una muestra de datos de la población objetivo y se obtiene información de esta muestra una única vez. También se tomará en cuenta que los datos se distribuyen aproximadamente como una distribución normal. Cuando la variable que nos interesa es un recuento o una proporción, los modelos para estos datos se esperaría tengan una distribución de Poisson o Binomial

La matriz de pesos se introduce en las especificaciones del modelo de regresión clásico con la finalidad de obtener la estructura de dependencia espacial del proceso económico, obteniendo así los modelos de regresión espaciales.

Hay una amplia tipología de modelos de regresión espacial. Dentro de esta amplia variedad de modelos veremos: el modelo de retardo espacial, modelo de error espacial, modelo de orden superior y el modelo espacial de Durbin.

Estos modelos parten del modelo de regresión lineal (múltiple), recordemos que tiene como forma funcional (Sánchez, 2018):

$$Y = \sum_{q=1}^{Q} X_q \beta_q + \varepsilon$$

Y su versión muestral

$$y_i = \sum_{q=1}^{Q} x_{iq} \beta_q + \varepsilon_i \quad i = 1, \cdots n$$

Donde:

- y_i= una observación de la variable dependiente (o de interés).
- x_{iq}=una observación en una variable explicativa con q=1,…,Q
- β_q= coeficiente de regresión que mide la influencia por si sola de la q-variable explicativa en la variable dependiente, es decir, mide el cambio

en Y por cada cambio unitario en X_q manteniendo las restantes variables explicativas constantes.

- ε_i= error aleatorio, Para estos terminos de error asumimos que ε_i son variables independientes e idénticamente distribuidas a una variable normal con media cero y varianza una constante σ^2.

En notación matricial este modelo se expresa como:

$$Y=X\beta + \varepsilon$$

Donde:

- Y= vector n x 1, n observaciones de la variable dependiente.
- X= matriz n x Q, muestra las observaciones de las variables explicativas.
- β= es el vector Q x 1 de parametros de regresión asociados a dichas variables explicativas.
- ε= vector de dimensión n x 1 de terminos de error.

Cuando estudiamos datos espaciales nos enfrentamos a que existe dependencia espacial, esta dependencia espacial puede estar presente en variables explicativas, variables dependientes o en los residuos. Cuando la dependencia espacial se encuentra en la variable dependiente los modelos se denominan modelos de retardo espacial mientras que si está en los residuos se denominan modelos de error espacial. Cuando está presente en las variables explicativas se llaman modelos de regresión cruzada o modelos espacialmente retardados, pero, en contraste con los otros dos modelos, no precisan de procedimientos especiales para la estimación.

3.4.1 Modelos de Retardo Espacial.

El modelo de retardo espacial básico, llamado modelo autorregresivo espacial de primer orden (SAR) son modelos de retardo espacial donde incluyen un retardo espacial como factor explicativo de la variable dependiente para formar la estructura de dependencia espacial del proceso, con la siguiente expresión (Sánchez, 2018):

$$Y=\rho WY + X\beta + \varepsilon; \quad \varepsilon \sim N(0, \sigma 2\, I)$$

Donde:

- Y= vector columna n x 1.
- X= matriz n x K que recoge una serie de variables exógenas.
- W= matriz de pesos espaciales.
- WY= es el retardo espacial de la variable Y.
- ρ= parámetro autorregresivo que determina tanto la intensidad como el carácter positivo o negativo de la dependencia espacial, viene definido por (w_{min}^{-1}, w_{max}^{-1}) donde w_{min} y w_{max} representa el máximo y mínimo auto valor de la matriz W.
- β= es el vector de parámetros K x 1.

- ε= corresponde con el termino de perturbaciones que se supone un ruido blanco, independiente e idénticamente distribuido.

Otra forma de expresar el modelo es en su forma reducida:

$$Y= (I – \rho W)^{-1} (X\beta + \varepsilon)$$

La esperanza del modelo en forma reducida es:

$$E(Y) = (I – \rho W)^{-1} E(X \beta)$$

El termino $(I – \rho W)^{-1}$ es el multiplicador espacial y nos dice que el valor esperado de cada y_i depende de la combinación lineal de X, obtenidas de las observaciones vecinas, escalado por el parámetro de dependencia ρ (Sánchez, 2018).

3.4.2 Modelos de Error Espacial.

El modelo de error espacial (SEM) es como una combinación de un modelo de regresión estándar con un modelo autorregresivo espacial en el término de error ε. Esto ocurre cuando la dependencia se da del proceso de error, ya que los errores de diferentes áreas pueden mostrar autocorrelación espacial (Morales-Oñate, 2022). Este modelo está dado por:

$$Y= X\beta + \upsilon$$
$$\upsilon= \lambda W \upsilon + \varepsilon$$

Donde,

- λ= parámetro autorregresivo del retardo espacial del error $W\upsilon$.
- ε= error aleatorio, se asume independiente e idénticamente distribuido.

En su forma estructural tenemos:

$$Y= X\beta + (I – \lambda W)^{-1} \varepsilon$$

Esta expresión conduce a un efecto de difusión espacial global, pero no hay un efecto multiplicador espacial (Morales-Oñate, 2022).

3.4.3 Modelo Espacial de Durbin.

El modelo espacial de Durbin (SDM) (Morales-Oñate, 2022) viene dado por su forma estructural:

$$y=\rho Wy + X\beta + WX\gamma + \varepsilon$$

Y su forma reducida:

$$y = (I - \rho W)^{-1} (X\beta + WX\gamma + \varepsilon)$$

El SDM incluye no solo la variable dependiente espacialmente rezagada y las variables explicativas, sino también las variables explicativas espacialmente rezagadas, WX. **Y** depende de factores de sí misma de la matriz X, más los mismos factores promediados en las n regiones vecinas (Morales-Oñate, 2022).

Capítulo 4

Aplicación

4.1 Sobre los Datos

Los datos para el siguiente análisis pertenecen a una aseguradora mexicana que desea mantener su nombre confidencial debido a la Ley Federal de Protección de Datos Personales. Estos datos fueron obtenidos de la base de siniestros reclamados a la asegurada desde el año 2008, donde se seleccionaron los siniestros reclamados por diabetes a las pólizas de Gastos Médicos Mayores del 1 de enero de 2008 al 31 de diciembre 2021, se utilizó corte al 31 de diciembre 2021 para estudiar años completos.

La aseguradora actualmente tiene oficinas en los estados de Yucatán, Quintana Roo, Campeche, Tabasco, Veracruz, Nuevo León, Jalisco, Guanajuato, Querétaro, Puebla y Ciudad de México. Inicialmente fue una aseguradora regional en el sureste de la república mexicana (Campeche, Yucatán y Quintana Roo), pero al ser comprada por una empresa nacional expandió sus sucursales desde el 2009.

El objetivo de utilizar estos datos es buscar si existe una relación espacial para los importes pagados por los siniestros reclamados y pagados por diabetes de acuerdo con el estado donde ocurre y buscar un modelo de regresión espacial que nos permita estimar cuanto se pagaría a un asegurado nuevo que pertenece a una póliza de gastos médicos mayores colectivo, que ya tiene el padecimiento de diabetes. Nuestro interés por estimar para asegurados que

pertenecen a pólizas de gastos médicos mayores colectivo y no para pólizas de gastos médicos mayores individual, es porque cuando se contrata una póliza nueva de gastos médicos colectivo, la empresa contratante brinda información de experiencia de siniestralidad, que incluye datos como sexo, edad, importe pagado, cuando inició su padecimiento, el nombre del padecimiento y fechas de todos los pagos generados por dicho padecimiento, en cambio para la individual no se entrega esa información.

La base de datos principalmente cuenta con las siguientes variables que nos ayudará para nuestro análisis:

Siniestro	Sexo	Estado donde Ocurre	Fecha de Nacimiento	Fecha Ocurrido	Fecha Movimiento	Pagos	Descripción del Padecimiento

El “siniestro” es el identificador de cada caso de diabetes; “sexo” es si es masculino o femenino; “estado donde ocurre” es el estado de la república mexicana donde ocurrió el siniestro; “fecha de nacimiento” es la del asegurado que reclamó el siniestro de diabetes; “fecha ocurrido” es la fecha de cuando inicia el asegurado con el padecimiento de diabetes; “fecha movimiento” es la fecha de cuando se pagaron gastos por el padecimiento; “pagos” es el importe pagado por cada gasto generado por el padecimiento; “descripción del padecimiento” es el nombre del padecimiento, que para nuestro caso todos son con nombre de diabetes. La base registra cada gasto pagado en pesos mexicanos en diferentes fechas desde que se apertura el siniestro, por lo que más adelante realizaremos agrupaciones que vayamos necesitando para nuestro análisis.

4.2 Análisis Exploratorio de los Datos

La base de datos cuenta con las variables de interés: pagos, sexo, fecha ocurrido y fecha movimiento, como a continuación podemos ver en la Tabla 2 que incluye algunos de los registros de los datos.

Siniestro	ado donde	SEXO	IA DE NACI	echa Ocur	cha Movim	Pagos	ión del Pa
5000535	YUCATAN	M	14/11/1992	2/7/2005	14/1/2008	1,602.92	DIABETES M
175	YUCATAN	F	25/6/1936	2/6/1999	15/1/2008	2,755.56	DIABETES M
5000535	YUCATAN	M	14/11/1992	2/7/2005	22/1/2008	683.72	DIABETES M
175	YUCATAN	F	25/6/1936	2/6/1999	27/1/2008	0.00	DIABETES M
3000001	Q. ROO	M	15/8/1945	16/11/2002	27/1/2008	0.00	DIABETES M
4000717	YUCATAN	F	29/5/1967	27/11/2004	27/1/2008	0.00	DIABETES M
5000506	Q. ROO	F	19/9/1936	19/7/2005	27/1/2008	0.00	DIABETES M
5000535	YUCATAN	M	14/11/1992	2/7/2005	27/1/2008	0.00	DIABETES M
5000535	YUCATAN	M	14/11/1992	2/7/2005	27/1/2008	0.00	DIABETES M
4000717	YUCATAN	F	29/5/1967	27/11/2004	28/1/2008	250.00	DIABETES M
3000001	Q. ROO	M	15/8/1945	16/11/2002	1/2/2008	2,703.67	DIABETES M
5000506	Q. ROO	F	19/9/1936	19/7/2005	1/2/2008	10,427.76	DIABETES M
5000535	YUCATAN	M	14/11/1992	2/7/2005	1/2/2008	672.23	DIABETES M
175	YUCATAN	F	25/6/1936	2/6/1999	8/2/2008	2,440.89	DIABETES M

Tabla 2. Extracto de la base de datos sobre los gastos generados y pagados por los siniestros de diabetes.

Antes de entrar en el análisis, agruparemos los datos por el número de siniestros. Como podemos ver en la Tabla 2, un número de siniestro tiene varios registros, que es cada vez que se realiza un pago, de esta forma conoceremos el importe total pagado hasta el 31 de diciembre de 2021 de cada siniestro, dónde ocurrió, sexo, cuándo comenzó y cuando fue su último pago. También crearemos dos variables, una para conocer cuántos años de antigüedad tiene cada siniestro y otra para la media pagada por antigüedad. Dado que conocemos las fechas de pago de cada registro podemos obtener la última fecha en la que se le pagó algún gasto. Realizando la diferencia de fecha última de pago y fecha de ocurrido obtendremos la antigüedad (“years”) de cada siniestro y sumando el total de pagos por siniestro y dividiendo entre los años de antigüedad tendremos la media (“payment”), dando como resultado la siguiente Tabla 3:

siniestro	estado	startDate	endDate	years	payment	sexo
5000535	YUCATAN	2005-07-02 00:00:00	2015-01-27 00:00:00	10	13748.947	M
175	YUCATAN	1999-06-02 00:00:00	2022-05-02 00:00:00	23	31795.8	F
3000001	Q. ROO	2002-11-16 00:00:00	2021-01-21 00:00:00	19	5850.04316	M
4000717	YUCATAN	2004-11-27 00:00:00	2008-09-22 00:00:00	4	187.5	F
5000506	Q. ROO	2005-07-19 00:00:00	2008-02-27 00:00:00	3	3475.92	F
5000457	YUCATAN	2005-08-05 00:00:00	2016-06-28 00:00:00	11	8555.79636	F
7000334	YUCATAN	2007-01-09 00:00:00	2022-06-16 00:00:00	16	5264.27438	M
8000423	YUCATAN	2008-02-05 00:00:00	2012-04-11 00:00:00	5	10221.422	M
8000500	YUCATAN	2008-02-27 00:00:00	2009-10-29 00:00:00	2	0	M
8000585	YUCATAN	2007-12-27 00:00:00	2012-09-26 00:00:00	5	4912.034	M
8000173	YUCATAN	2007-08-31 00:00:00	2015-09-09 00:00:00	8	4036.5825	M
8001190	YUCATAN	2008-06-19 00:00:00	2010-08-06 00:00:00	3	4149.46	F
8001551	YUCATAN	2008-05-16 00:00:00	2022-05-20 00:00:00	14	21628.5857	F
8001587	YUCATAN	2008-08-20 00:00:00	2013-08-23 00:00:00	5	44117.244	M

Tabla 3. Importe total pagado por siniestro.

Terminando la agrupación nos encontramos con 435 siniestros de diabetes reclamados a la compañía de seguros desde 01 de enero de 2008 al 31 de diciembre de 2021, identificando donde ocurrieron, la antigüedad del siniestro, sexo y el importe pagado total.

Para el análisis espacial nos interesa ver cómo se comporta la variable importe pagado por cada estado, para esto usaremos la media pagada por año antigüedad de los siniestros de cada estado y analizando las dos categorías de sexo, femenino y masculino, Tabla 4:

estado	pay_femenino	pay_masculino	count_femenino	count_masculino
AGUASCALIE	$ 31,462.71	$ 18,797.90	1	2
BAJA CALIFO	$ 14,667.57	$ 311.09	1	2
CDMX	$ 30,092.97	$ 56,630.96	59	76
CHIAPAS	$ -	$ -	0	1
CHIHUAHUA	$ -	$ 2,643.38	0	1
COAHUILA	$ 49,705.49	$ 19,641.00	5	6
DURANGO	$ -	$ -	0	1
EXTRANJERO	$ -	$ -	1	0
GUANAJUAT	$ 7,110.81	$ 144.72	5	2
HIDALGO	$ -	$ 13,515.09	0	1
JALISCO	$ 21,523.08	$ 27,388.66	15	12
MEXICO	$ 73,029.77	$ 34,937.50	13	21
MICHOACAN	$ -	$ 1,338.15	0	1
MORELOS	$ 41,919.13	$ 5,590.38	1	1
NUEVO LEON	$ 62,565.94	$ 47,930.56	22	45
PUEBLA	$ 11,561.62	$ 32,999.11	7	11
Q. ROO	$ 7,043.15	$ 13,248.53	7	10
QUERETARO	$ 7,457.44	$ 19,714.06	4	4

Tabla 4. Importe pagado medio por estado y por sexo.

A modo de ilustración, en la siguiente imagen (Figura 9) veremos la representación de los pagos para hombres y mujeres en la república mexicana.

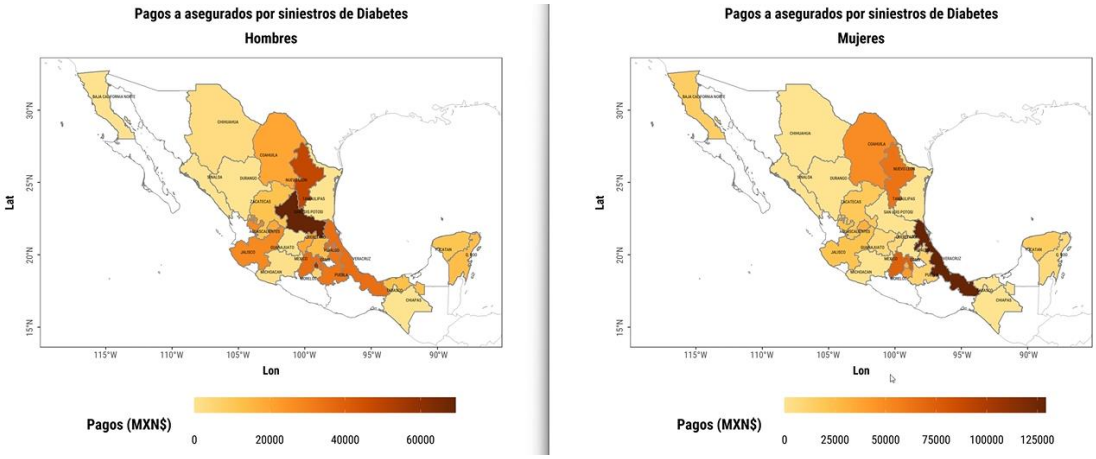


Figura 9 : Elaboración propia, media pagada por estado y sexo.

Dado que los importes en mujeres son más grandes que los hombres, procedemos a estandarizar los datos (Figura 10):

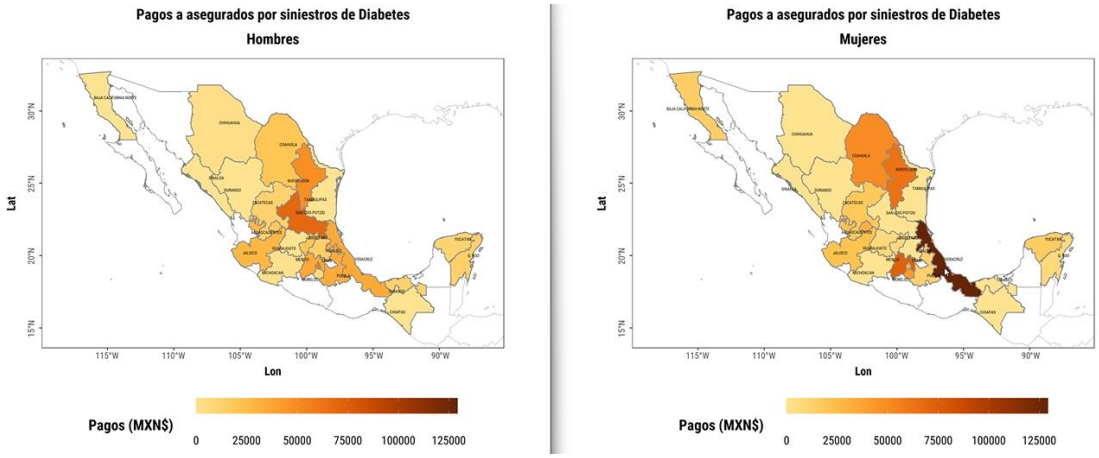


Figura 10: Elaboración propia, media pagada por estado y sexo, estandarizada.

Observamos que en el estado de chiapas el importe pagado medio fue menos tanto para hombres y mujeres en relación a otros estados, en cambio para las mujeres, en el estado de Veracruz es donde mas se les pagó y en el caso de los hombres fue en San Luis Potosi. Los estados en blanco indican que no se presentaron casos de siniestros de diabetes donde la aseguradora hubiera pagado gastos, esto igual se debe a que la compañía de seguros no cuenta con oficina en esos estados lo que provoca que se tenga pocos asegurados en esas zonas o ningún asegurado.

En relación a la duración de los siniestros de diabetes, obtenemos el siguiente mapa (Figura 11) donde se representa la cantidad de meses promedio por estado y sexo.

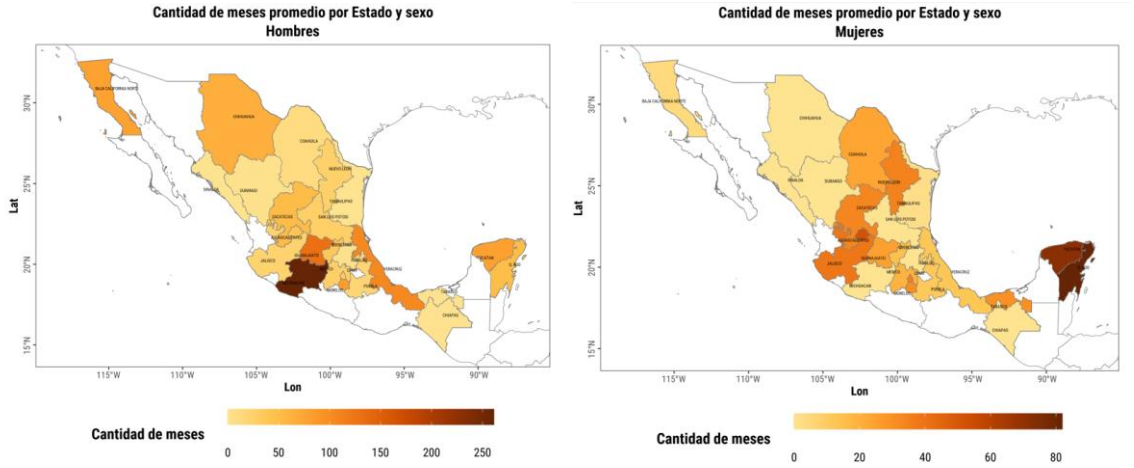


Figura 11: Elaboración propia, tiempo medio en meses por estado y sexo de los siniestros de diabetes.

Con el anterior mapa podemos observar que para los hombres los siniestros de diabetes que ocurren en el estado de Michoacan, en promedio duran 260 meses (casi 22 años) y para mujeres los siniestros que ocurren en el estado de Quintana Roo duran en promedio 80 meses (casi 7 años), otros estados como Chiapas, Durango, Sinaloa y Tamaulipas presentan la menor duración tanto para hombres como para mujeres.

4.3 Correlación Espacial

Empezaremos analizando si en nuestros datos existe autocorrelación espacial, para esto hemos utilizado una matriz de pesos tipo reina para calcular la medida local del estadístico de Moran (Indicador LISA), ya que nos ayudará a evaluar cada estado de la república mexicana para conocer si sus importes pagados por siniestros de diabetes son altos o bajos con relación a sus estados vecinos y de igual forma para la duración.

De esta forma obtenemos la siguiente clasificación espacial (Figura 12) según el índice de LISA para Hombres y Mujeres.

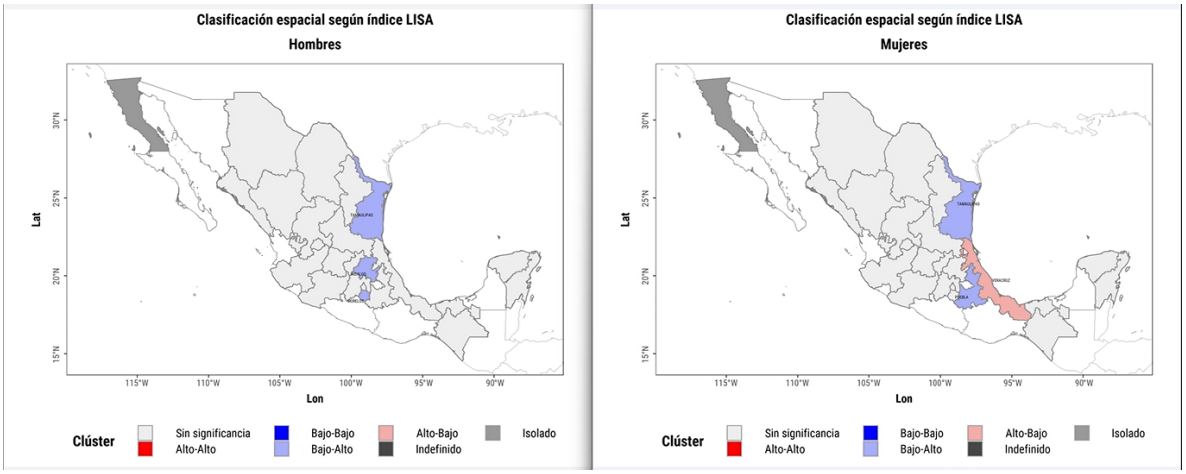


Figura 12: Elaboración propia, clasificación espacial según indice de LISA de los pagos de siniestros de diabetes.

A la vista del gráfico anterior se tiene la siguiente información distinguida por sexos.

- Hombres: correlación alta que se da en los estados de Tamaulipas, Hidalgo y Morelos, estos tres estados se caracterizan porque tienen unos pagos bajos, mientras que los estados con los que limitan se dan sus pagos altos.
- Mujeres: sucede lo mismo, solo que, con Tamaulipas y Puebla, para el caso de Veracruz la correlación es que en ese estado la media de pagos es alta, pero a su alrededor es bajo.

Para el resto de los estados nos encontramos con que no existe suficiente evidencia estadística para afirmar que existe correlación espacial de los estados en colores grises y su alrededor, es decir, no están correlacionados entre sí y los estados en blanco lo mismo, ya que no se tienen casos de siniestros de diabetes en ellos.

Para la duración obtuvimos la siguiente clasificación espacial según índice LISA.

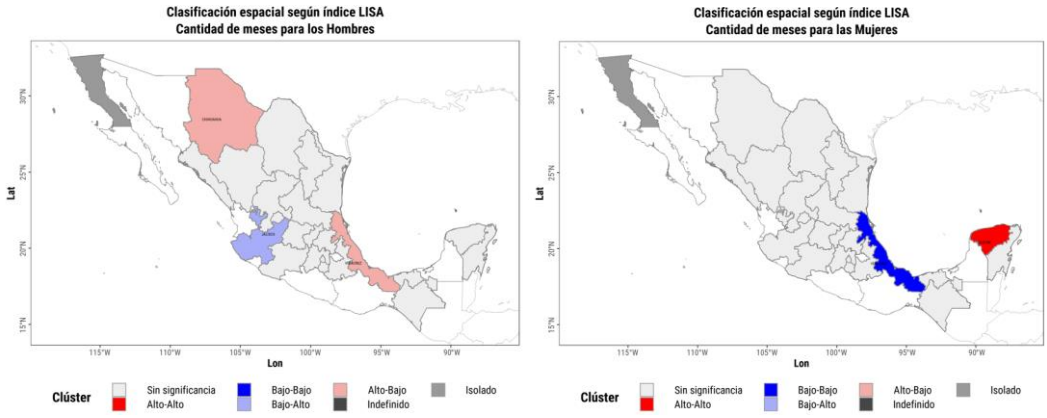


Figura 13: Elaboración propia, clasificación espacial según indice LISA de la duración media en meses.

A la vista del gráfico anterior se tiene la siguiente información distinguida por sexos.

- Hombres: en el estado de Chihuahua se presenta un resultado alto-bajo, lo que significa que en él se presenta una duración del siniestro de diabetes alto y los estados vecinos Coahuila, Durango y Sinaloa tienen un comportamiento bajo en la cantidad de meses promedio que dura el siniestro. Veracruz, del mismo modo, presenta una correlación alto-bajo, donde los estados vecinos con baja duración son Tabasco, Chiapas, Puebla, Hidalgo, San Luis Potosí y Tamaulipas.
- Mujeres: se observa para Yucatán una relación Alto-Alto, donde su duración es alta al igual que su estado vecino Quintana Roo. En Veracruz se tiene Bajo-Bajo, su duración es baja al igual que sus estados vecinos Tabasco, Chiapas, Puebla, Hidalgo, San Luis Potosí y Tamaulipas

Para el resto de los estados nos encontramos con que no existe suficiente evidencia estadística para afirmar que existe correlación espacial de los estados en colores grises y su alrededor, es decir, no están correlacionados entre sí y los estados en blanco lo mismo, ya que no se tiene casos de siniestros de diabetes en ellos.

4.4 Modelos Espaciales

Primero realizamos un Modelo de regresión clásico sin efectos espaciales, donde se estimará el pago medio de las mujeres y hombres en función de los meses de duración media y la cantidad de siniestros, el resultado de este modelo es el siguiente:

Mujeres:

```
Call:
lm(formula = pay_femenino ~ months_femenino + count_femenino,
    data = tble)

Residuals:
   Min     1Q Median     3Q    Max
-22905 -16904 -13329   -511 110174

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      16903.6     9072.8   1.863   0.0765 .
months_femenino    141.8      337.1   0.421   0.6783
count_femenino     205.1      519.8   0.395   0.6972
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32030 on 21 degrees of freedom
Multiple R-squared:  0.03147,  Adjusted R-squared:  -0.06077
F-statistic: 0.3411 on 2 and 21 DF,  p-value: 0.7148
```

Encontrando que, no hay correlación ya que el p-valor de “months_femenino” y “count_femenino” es muy alto, por lo que no hay significancia.

Hombres:

```
Call:
lm(formula = pay_masculino ~ months_masculino + count_masculino,
    data = tble)

Residuals:
   Min     1Q Median     3Q    Max
-22026  -9369  -1822   5920  53995

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      14284.66    4936.21   2.894  0.00868 **
months_masculino   -52.62      59.44  -0.885  0.38605
count_masculino    594.19     189.03   3.143  0.00491 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16400 on 21 degrees of freedom
Multiple R-squared:  0.3478,   Adjusted R-squared:  0.2856
F-statistic: 5.598 on 2 and 21 DF,  p-value: 0.01125
```

Donde obtenemos que, hay una relación con la cantidad de siniestros ya que el p-valor de “count_masculino” es .00491, cercano a cero, resultando que por cada nuevo siniestro de diabetes se tiene un aumento de $594.19 pesos

Podemos eliminar la variable “months_masculino” por no ser significativa:

```
Call:
lm(formula = pay_masculino ~ count_masculino, data = tble)

Residuals:
   Min     1Q Median     3Q    Max
-24044 -11607  -2808   6603  55171

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      11607.1     3881.8   2.990  0.00675 **
count_masculino    607.9      187.5   3.243  0.00373 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16320 on 22 degrees of freedom
Multiple R-squared:  0.3234,   Adjusted R-squared:  0.2927
F-statistic: 10.52 on 1 and 22 DF,  p-value: 0.003734
```

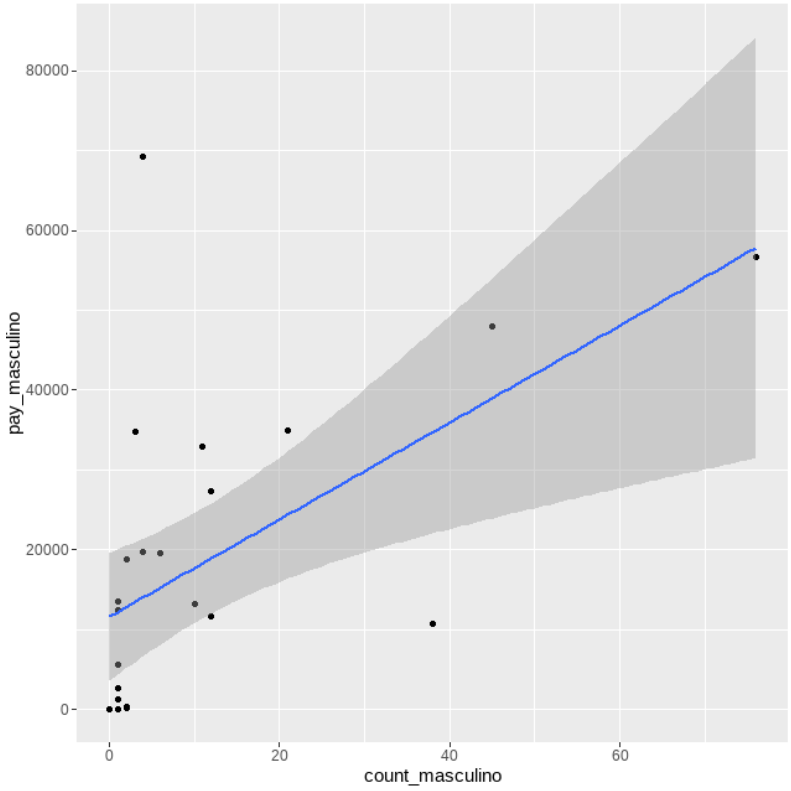


Figura 14: Elaboración propia, grafico de dispersión de cantidad de siniestros de los hombres vs la media pagada hombres.

Resultando el modelo de regresión lineal simple: Y=607.9x + 11,607.1, notemos que este modelo está evaluando todos los estados de la república, por lo que nos encontramos un valor cero en la cantidad de siniestros de hombres debido a que en el estado de Tamaulipas no se presentaron siniestros de diabetes.

Cuando analizamos la correlación espacial, nos encontramos que muy pocos estados tenían una relación del nivel de pagos con sus estados vecinos, además que nuestra base de datos contiene información de tan solo 20

polígonos, 20 estados de la república en el que se cuenta información, estos no colindan todos entre sí, por lo que son datos insuficientes para realizar un modelo espacial, en virtud de eso se realizó un modelo lineal para todos y cada uno de los estados.

A continuación, podemos visualizar en el siguiente mapa (Figura 15) donde describe la cantidad de siniestros por estados:

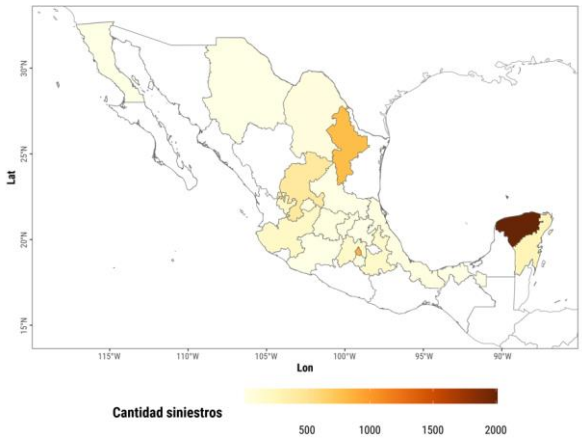


Figura 15: Elaboración propia, mapa de frecuencia por estado de la republica mexicana.

Se pretende estimar un modelo lineal para cada estado de la república mexicana con la finalidad de buscar los estados donde sí existe o no una relación en la duración del siniestro con el importe pagado total por siniestro.

Lo anterior se calcula con la Tabla 5 siguiente:

siniestro	estado	duracion_m	sexo	pago
5000535	YUCATAN	114	M	$ 137,489.47
175	YUCATAN	275	F	$ 710,042.78
4000717	YUCATAN	45	F	$ 750.00
5000457	YUCATAN	130	F	$ 94,113.76
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
3000001	Q. ROO	218	M	$ 111,150.82
5000506	Q. ROO	30	F	$ 10,427.76
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
7000262	VERACRUZ	140	M	$ 50,350.65
20001144	VERACRUZ	162	M	$ 2,611.20
21004119	VERACRUZ	9	F	$ 72,498.54
21009433	VERACRUZ	1	M	$ 100,000.00
9001167	CDMX	1	M	$ 33,358.72
9001484	CDMX	1	M	$ 23,007.51
10000958	CDMX	83	F	$ 95,857.49

Tabla 5: Elaboración propia, extracto de la base de datos del total pagado y duración por siniestro y sexo.

Obteniendo el siguiente resultado:

estadistico	estimado	stdError	t_value	p_value	sexo	estado
(Intercept)	82652.3977	31467.9412	2.6265588	0.01363368	Masculino	YUCATAN
duracion_m	-75.1586337	269.764674	-0.27860814	0.78252263	Masculino	YUCATAN
(Intercept)	-14265.7236	27209.5256	-0.52429152	0.60304313	Femenino	YUCATAN
duracion_m	**1514.58338**	**273.629209**	**5.53516705**	**2.2885E-06**	**Femenino**	**YUCATAN**
(Intercept)	27351.1224	28346.2736	0.96489305	0.36674108	Masculino	Q. ROO
duracion_m	558.54438	344.458971	1.62151207	0.14893848	Masculino	Q. ROO
(Intercept)	-85680.4166	111918.21	-0.76556278	0.48661297	Femenino	Q. ROO
duracion_m	1860.05684	1140.00642	1.63161962	0.17809531	Femenino	Q. ROO
(Intercept)	103183.923	25957.727	3.97507545	0.15689715	Masculino	VERACRUZ
duracion_m	-516.798416	209.981628	-2.46116015	0.24569466	Masculino	VERACRUZ
(Intercept)	72498.54	NA	NA	NA	Femenino	VERACRUZ
(Intercept)	117771.178	19576.7477	6.01587042	7.9771E-08	Masculino	CDMX
duracion_m	210.495708	326.345829	0.64500811	0.52109221	Masculino	CDMX
(Intercept)	62431.1096	15797.6846	3.95191519	0.00023928	Femenino	CDMX
duracion_m	223.243136	274.762921	0.81249368	0.42028497	Femenino	CDMX
(Intercept)	53413.2589	18197.079	2.93526555	0.01490581	Masculino	JALISCO
duracion_m	-308.92537	669.65472	-0.46132038	0.65443995	Masculino	JALISCO
(Intercept)	68178.9583	33353.6287	2.04412416	0.06352705	Femenino	JALISCO
duracion_m	-67.4174916	606.358028	-0.1111843	0.91330822	Femenino	JALISCO
(Intercept)	-26008.8808	24687.0496	-1.05354351	0.31688234	Masculino	ZACATECAS
duracion_m	**1880.50637**	**367.958298**	**5.11065078**	**0.00045687**	**Masculino**	**ZACATECAS**
(Intercept)	37340.1215	34126.0787	1.09418143	0.31584495	Femenino	ZACATECAS
duracion_m	325.657819	809.548131	0.4022711	0.70142358	Femenino	ZACATECAS
(Intercept)	59558.0934	17115.002	3.47987652	0.00309221	Masculino	MEXICO
duracion_m	391.761916	394.646459	0.99269082	0.33563703	Masculino	MEXICO
(Intercept)	140682.76	57495.9124	2.44683064	0.03444407	Femenino	MEXICO
duracion_m	-1156.28162	2732.22725	-0.42320111	0.68110576	Femenino	MEXICO
(Intercept)	86629.7572	18533.7522	4.67416183	3.3234E-05	Masculino	NUEVO LEON
duracion_m	86.5919526	447.579807	0.19346707	0.84757228	Masculino	NUEVO LEON
(Intercept)	147186.712	99247.2488	1.48303065	0.15536421	Femenino	NUEVO LEON
duracion_m	1455.02715	1892.08921	0.76900557	0.45186322	Femenino	NUEVO LEON
(Intercept)	59240.99	NA	NA	NA	Masculino	AGUASCALIENTES
(Intercept)	157313.56	NA	NA	NA	Femenino	AGUASCALIENTES
(Intercept)	12915.6707	6331.73774	2.03983033	0.11096502	Masculino	COAHUILA
duracion_m	**1681.60418**	**378.844979**	**4.43876591**	**0.01134478**	**Masculino**	**COAHUILA**
(Intercept)	134356.44	61915.175	2.17000824	0.11845517	Femenino	COAHUILA
duracion_m	-5101.70885	4431.56576	-1.15122039	0.33306984	Femenino	COAHUILA
(Intercept)	1866.51	NA	NA	NA	Masculino	BAJA CALIFORNIA NORTE
(Intercept)	14667.57	NA	NA	NA	Femenino	BAJA CALIFORNIA NORTE
(Intercept)	33700.1462	91177.9051	0.36960869	0.73041	Masculino	PUEBLA
duracion_m	2984.6094	2564.91717	1.16362799	0.30925405	Masculino	PUEBLA
(Intercept)	10524.5218	6788.56619	1.55033058	0.21884418	Femenino	PUEBLA
duracion_m	**1893.49602**	**247.817313**	**7.64069303**	**0.004655**	**Femenino**	**PUEBLA**
(Intercept)	5788.8	NA	NA	NA	Masculino	GUANAJUATO
(Intercept)	18026.5982	11267.1521	1.59992499	0.25074676	Femenino	GUANAJUATO
duracion_m	122.644478	307.165408	0.39927829	0.72828925	Femenino	GUANAJUATO
(Intercept)	39132.63	NA	NA	NA	Masculino	MORELOS
(Intercept)	90315.81	NA	NA	NA	Femenino	MORELOS
(Intercept)	55064.268	43064.645	1.2786421	0.32934181	Masculino	QUERETARO
duracion_m	-479.625847	903.874102	-0.53063347	0.64870048	Masculino	QUERETARO
(Intercept)	23308.1195	24320.3787	0.95837815	0.51352819	Femenino	QUERETARO
duracion_m	-270.922368	1448.25747	-0.18706782	0.88226964	Femenino	QUERETARO

Tabla 6: Elaboración propia, parámetros modelo lineal por estado y sexo.

Los siguientes estados presentan una colinealidad entre la cantidad de pago y la cantidad de meses mientras duró el siniestro, a continuación, el gráfico nos permite identificar el comportamiento de estas dos variables, según sea el sexo y por estado (ver Figura 16).

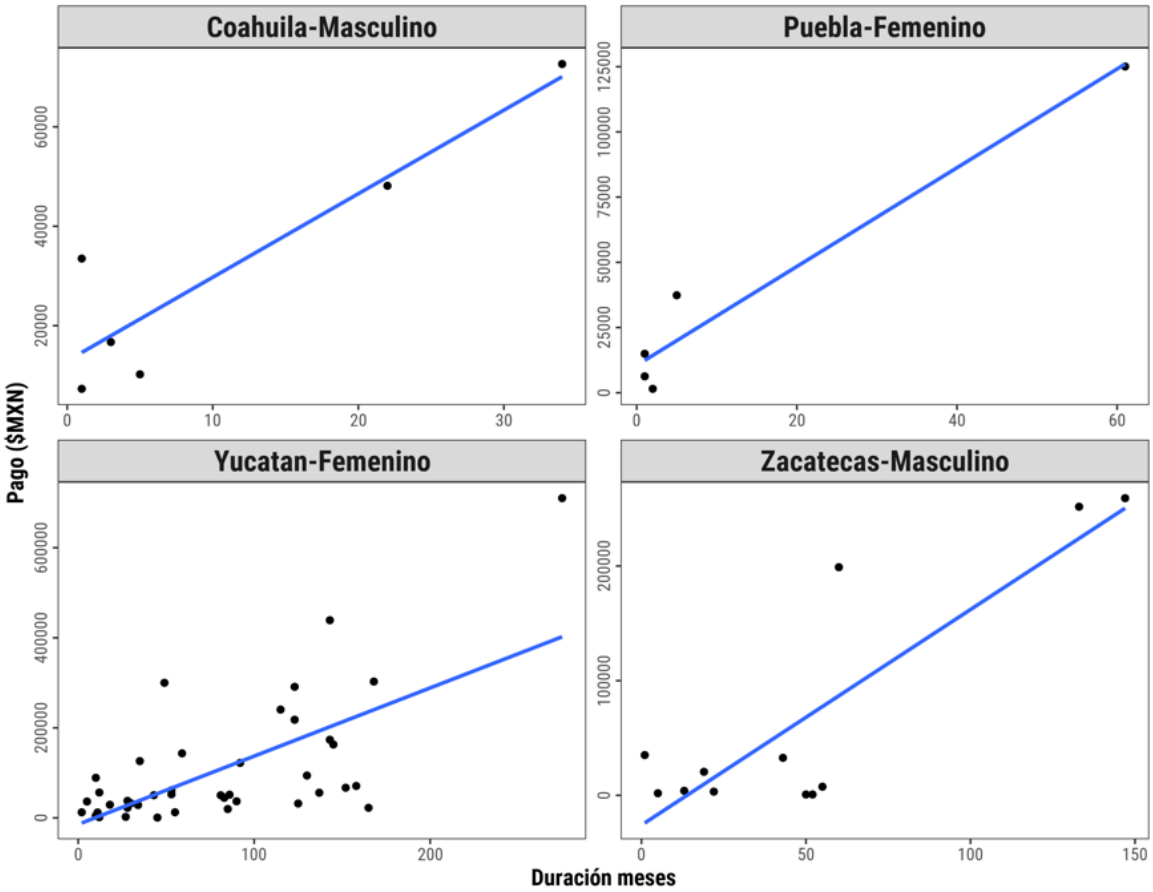


Figura 16: Elaboración propia, correlación de los estados-sexo con significancia

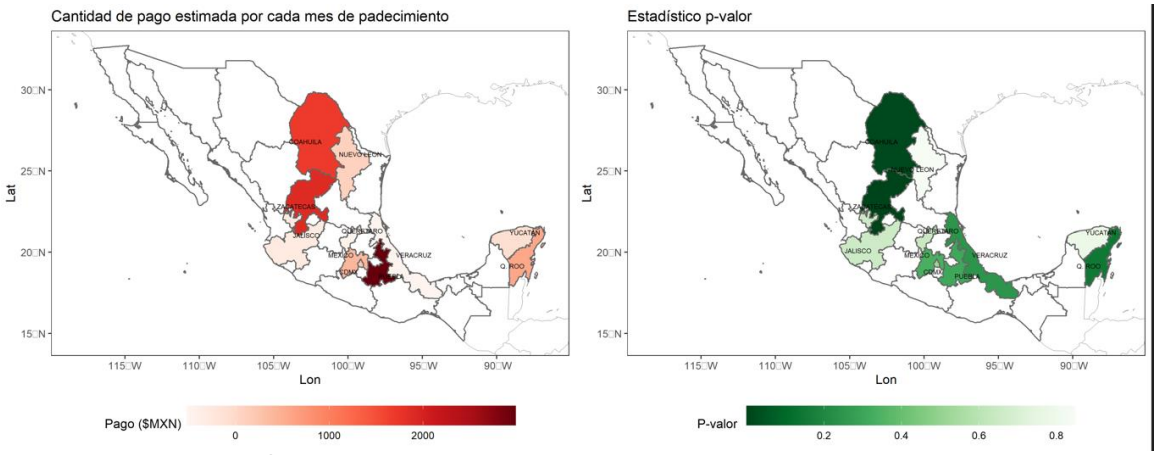


Figura 17: Elaboración propia, contraste de los pagos de los hombres por estado con su p-valor por estado.

La Figura 17 nos dice que para los hombres donde el siniestro ocurrió en Coahuila y Zacatecas por cada nuevo mes se estima pagar mas de 2 mil pesos.

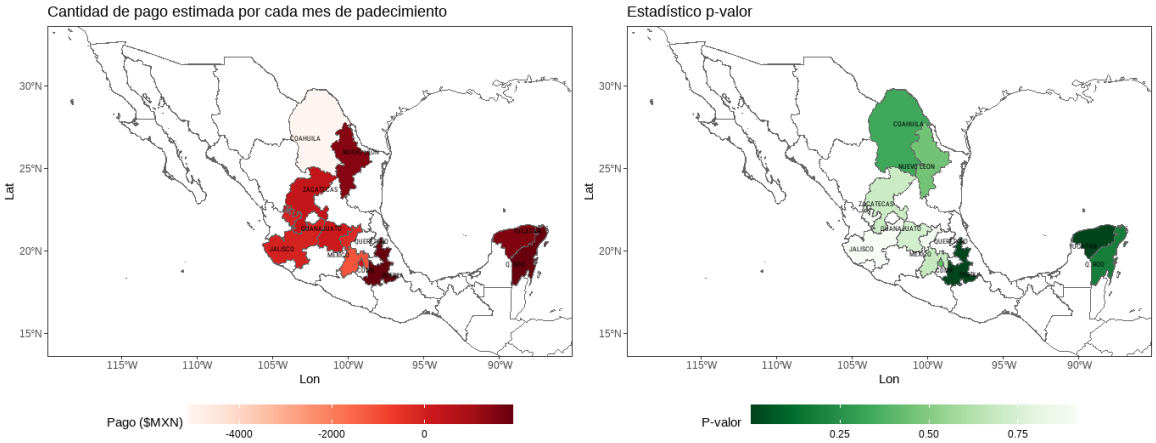


Figura 18: Elaboración propia, contraste de los pagos de los mujeres por estado con su p-valor por estado.

La figura 18 nos dice que para las mujeres donde el siniestro ocurrió en Yucatán y Puebla por cada nuevo mes se estima pagar alrededor de 500 a mil pesos.

Veamos qué resultado tendríamos estimando el modelo de Retardo Espacial:

Hombres:

```
Call:lagsarlm(formula = f1, data = shpf_hmbr, listw = seaw, zero.policy
= TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-54789.1 -18210.8  -3506.7  16661.9  59232.0

Type: lag
Regions with no neighbours included:
 2
Coefficients: (numerical Hessian approximate standard errors)
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) 65677.00   17314.76  3.7931 0.0001488
duracion_m   -265.33     149.60 -1.7736 0.0761302

Rho: 0.12764, LR test value: 0.29889, p-value: 0.58458
Approximate (numerical Hessian) standard error: 0.23028
    z-value: 0.55431, p-value: 0.57937
Wald statistic: 0.30726, p-value: 0.57937

Log likelihood: -176.3907 for lag model
ML residual variance (sigma squared): 951820000, (sigma: 30852)
Number of observations: 15
Number of parameters estimated: 4
AIC: 360.78, (AIC for lm: 359.08)
```

Rho es el indice de dependencia espacial, que para los hombres fue de .12 y con un p-valor .58, lo que nos indica que no existe una correlación espacial global entre la cantidad de meses y la cantidad de dinero pagado medio de los hombres.

Mujeres:

```
Call:lagsarlm(formula = f1, data = shpf_mjrs, listw = seaw, zero.policy
= TRUE)

Residuals:
   Min     1Q Median     3Q    Max
-57691 -34030  -8029  20825 116733

Type: lag
Regions with no neighbours included:
 2
Coefficients: (numerical Hessian approximate standard errors)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 60752.26   26368.03  2.3040  0.02122
duracion_m    683.29     515.43  1.3257  0.18495

Rho: -0.042397, LR test value: 0.021913, p-value: 0.88232
Approximate (numerical Hessian) standard error: 0.28519
    z-value: -0.14866, p-value: 0.88182
Wald statistic: 0.0221, p-value: 0.88182

Log likelihood: -182.7731 for lag model
ML residual variance (sigma squared): 2243100000, (sigma: 47361)
Number of observations: 15
Number of parameters estimated: 4
AIC: 373.55, (AIC for lm: 371.57)
```

De igual forma obtuvimos un p-valor alto de .88, indicando que no existe autocorrelación espacial global entre la cantidad de meses y la cantidad de dinero pagado medio de las mujeres.

Debido a que no existe correlación espacial global nos topamos con que los modelos de regresión espacial no serían significativos como vimos en el modelo de retardo espacial, por lo que no se continua estimando más modelos espaciales, incluso al realizar el análisis de regresión lineal simple para todos los siniestros de cada estado, únicamente encontramos significancia para 4 estados, por lo que para nuestro caso práctico se deberia buscar otras alternativas para estimar el pago de siniestros de diabetes.

Conclusiones

Una vez realizado el análisis espacial con la información de diabetes de una aseguradora, nos permitió explorar los datos mediante ilustraciones de mapas que nos permitieron identificar los estados de la república mexicana en que se pagaba una media mayor o menor, tanto para hombres como mujeres. En este sentido, la media pagada mayor se genera en las mujeres en el estado de Veracruz. El promedio de duración del padecimiento no es proporcional al importe medio pagado, por ejemplo, en Michoacán se encontró un mayor promedio de duración en el caso de los hombres, pero no fue el estado que presentó la mayor media pagada. En cambio, San Luis Potosí fue quien presentó el importe medio pagado mayor y su promedio de duración estaba entre los estados con baja duración.

Después continuamos con determinar si existe autocorrelación espacial. Para esto nos ayudamos con el índice de LISA, donde se encontraron con pocos estados que tenían autocorrelación espacial, tanto para la media pagada y el promedio de duración en hombres y mujeres. La mayoría de los estados de la república mexicana no tenían correlación espacial entre sí. Dado el resultado anterior, nos enfrentamos a que no podríamos encontrar un modelo de regresión espacial debido a la falta de dependencia espacial, esto nos llevó a manera de exploración, estimar un modelo lineal para cada estado, resultando significancia únicamente para dos estados para hombres y dos estados para mujeres.

Si bien en este trabajo no se pudo encontrar alguna relación del tipo de regresión espacial o incluso lineal para todos los estados, un suscriptor de seguros podrá encontrar información importante para tomar posturas conservadoras o agresivas al cotizar un negocio.

Bibliografía

AMIS. (s.f.). *Asociación Mexicana de Instituciones de Seguros*. Obtenido de https://sitio.amis.com.mx/comites/accidentes-y-enfermedades/#1507513104236-97e90cfd-63ae

Anselin, L. (2001). *Spatial econometrics. En A companion to theoretical* (págs. 310-330). Baltagi, Oxford: Basil Blackwell;.

Aroca, P. (2000). *Econometría espacial: Una Herramienta Para el Análisis de la Economía Regional.* IDEAR-Universidad Católica del Norte.

Cressie, N. (1993). *Statistics for spatial data.* New York, Wiley.

Damodar N. Gujarati, D. C. (2009). *Econometría* (Quinta ed.). McGraw-Hill.

David E. Rodriguez Guevara, G. J. (2017). *Principios de Econometría.* Fodo Editorial ITM.

Díaz, M. F. (2018). *Estadística descriptiva para distribución espacial.* Universidad Alberto Hurtado.

Hildegart, M. F. (2018). *Una Nueva econometría. Automatización, big data, econometria espacial y estructural.* Bahía Blanca: Editorial de la Universidad Nacional.

INEGI. (2021). *Instituto Nacional de Estadistica y Geografía.* Obtenido de https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2021/EAP_Diabetes2021.pdf

Luis Quintana Romero, M. A. (2017). *Econometría Aplicada utilizando R.* Universidad Nacional Autónoma de México.

Morales-Oñate, V. (2022). *Econometría Espacial*. Obtenido de https://bookdown.org/victor_morales/SpatialEconometrics/regresi%C3%B3n-lineal.html#modelo-con-error-espacial-sem

Moran, P. (1950). *Notes on Continuous Stochastic Phenomena* (Vol. 37). Biometrika.

Organización Mundial de la Salud. (10 de Noviembre de 2021). *Diabetes.* Obtenido de https://www.who.int/es/news-room/fact-sheets/detail/diabetes

Sánchez, J. A. (2018). *Modelos de Regresión para datos espaciales.* (Trabajo Fin de Grado). Universidad de Sevilla.

SESA. (2020). *Centro Estadistico AMIS.* Obtenido de Centro Estadístico del Sector Asegurador: https://centroestadisticoamis.mx/accidentes-y-enfermedades-reportes-ejecutivos/

Tobler, W. (1970). *A computer model simulation of urban growth in the.* Economic Geography 46 (2).

Yrigoyen, C. C. (Abril de 2003). *Econometría Espacial Aplicada a la Predicción-Extrapolación de datos microterritoraiales.* (Tesis Doctoral). Madrid: Consejería de Economía e Innovación Tecnológica.

Anexos

Sintaxis utilizada en el Software R para los gráficos de la aplicación del trabajo.

1. Figura 10: Elaboración propia, media pagada por estado y sexo, estandarizada.

```
# Load libraries ---------------------------------------------------------
require(pacman)
pacman::p_load(terra, showtext, extrafont, ggrepel, lubridate, ggspatial,
               cowplot, ggpubr, cptcity, rnaturalearthdata, rnaturalearth,
               RColorBrewer, openxlsx, fs, sf, readxl, tidyverse, gtools, rgeos,
               stringr, rgeoda, glue)

g <- gc(reset = T); rm(list = ls()); options(scipen = 999)

# Fonts ------------------------------------------------------------------
font_add_google(family = 'Roboto', name = 'Roboto Condensed')
showtext_auto()

# Load data --------------------------------------------------------------
shpf <- st_read('gpkg/summary_pagos_promedio_estado_sexo_v2.gpkg')
mex1 <- st_read('shp/mex1.shp')
wrld <- ne_countries(returnclass = 'sf', scale = 50) %>% filter(region_un ==
                                                                  'Americas')

# To make the maps -------------------------------------------------------

# Pagos ------------------------------------------------------------------
pays <- shpf %>% dplyr::select(ENTIDAD, starts_with('pay'))
pays <- pays %>% gather(sexo, pago, -ENTIDAD, -geom)
pays <- mutate(pays, sexo = ifelse(sexo == 'pay_femenino', 'Mujeres',
                                   'Hombres'))

st_write(pays, 'gpkg/summary_pagos_promedio_estado_sexo_2.gpkg')

# Months -----------------------------------------------------------------
mnts <- shpf %>% dplyr::select(ENTIDAD, starts_with('months'))
mnts <- mnts %>% gather(sexo, months, -ENTIDAD, -geom)
mnts <- mutate(mnts, sexo = ifelse(sexo == 'months_femenino', 'Meses Mujeres',
                                   'Meses Hombres'))
mnts
```

```
# Coordinates -------------------------------------------------------------
crds <- mutate(as.data.frame(st_coordinates(st_centroid(pays))),
               ENTIDAD = pays$ENTIDAD)
crds <- distinct(crds)

# -------------------------------------------------------------------------
# Mapas pagos -------------------------------------------------------------
# -------------------------------------------------------------------------

# Mujeres -----------------------------------------------------------------

g_muj <- ggplot() +
  geom_sf(data = filter(pays, sexo == 'Mujeres'), aes(fill = pago), col =
            'grey60', lwd = 0.5)  +
  scale_fill_gradientn(colors = brewer.pal(n = 9, name = 'YlOrBr')[3:9],
                       limits = c(0, 129000)) +
  geom_sf(data = mex1, fill = NA, col = 'grey50', lwd = 0.3) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +

  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD), family =
                    'Roboto', size = 5) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  labs(x = 'Lon', y = 'Lat', fill = 'Pagos (MXN$)') +
  ggtitle(label = glue('Pagos a asegurados por siniestros de Diabetes'),
          subtitle = glue('Mujeres')) +
  theme_bw() +
  theme(legend.position = 'bottom',
        legend.key.width = unit(4, 'line'),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(family = 'Roboto', size = 50, face = 'bold',
                                  hjust = 0.5),
        plot.subtitle = element_text(family = 'Roboto', size = 50, face =
                                       'bold', hjust = 0.5),
        axis.text.y = element_text(angle = 90, hjust = 0.5, family = 'Roboto',
```

```
                                    size = 50),
        legend.text = element_text(family = 'Roboto', size = 40))
##para ver el mapa aqui en R
g_muj

##Guardar imagen del mapa
ggsave(plot  = g_muj, filename = './png/maps/average_pays/pago_mujeres.png',
       units = 'in', width = 9, height = 7, dpi = 300)

# Hombres ----------------------------------------------------------------

g_hmb <- ggplot() +
  geom_sf(data = filter(pays, sexo == 'Hombres'), aes(fill = pago), col =
            'grey60', lwd = 0.5)  +
  scale_fill_gradientn(colors = brewer.pal(n = 9, name = 'YlOrBr')[3:9],
                       limits = c(0, 129000)) +
  geom_sf(data = mex1, fill = NA, col = 'grey50', lwd = 0.3) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +

  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD), family =
                    'Roboto', size = 5) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  labs(x = 'Lon', y = 'Lat', fill = 'Pagos (MXN$)') +
  ggtitle(label = glue('Pagos a asegurados por siniestros de Diabetes'),
          subtitle = glue('Hombres')) +
  theme_bw() +
  theme(legend.position = 'bottom',
        legend.key.width = unit(4, 'line'),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(family = 'Roboto', size = 50, face = 'bold',
                                  hjust = 0.5),
        plot.subtitle = element_text(family = 'Roboto', size = 50, face =
                                       'bold', hjust = 0.5),
        axis.text.y = element_text(angle = 90, hjust = 0.5, family = 'Roboto',
                                   size = 30),
        axis.text.x = element_text(family = 'Roboto', size = 30),
        axis.title = element_text(family = 'Roboto', size = 40, face = 'bold'),
        legend.title = element_text(family = 'Roboto', face = 'bold', size =50),
        legend.text = element_text(family = 'Roboto', size = 40))
#Ver mapa en R
g_hmb
#Guardar Mapa
ggsave(plot  = g_hmb, filename = './png/maps/average_pays/pago_hombres.png',
       units = 'in', width = 9, height = 7, dpi = 300)
```

2. Figura 11: Elaboración propia, tiempo medio en meses por estado y sexo de los siniestros de diabetes.

```
g_muj_mnt <- ggplot() +
  geom_sf(data = filter(mnts, sexo == 'Meses Mujeres'), aes(fill = months),
          col = 'grey60', lwd = 0.5)  +
  scale_fill_gradientn(colors = brewer.pal(n = 9, name = 'YlOrBr')[3:9]) + # ,
  limits = c(0, 261)
  geom_sf(data = mex1, fill = NA, col = 'grey50', lwd = 0.3) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +

  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD), family =
                    'Roboto', size = 5) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  labs(x = 'Lon', y = 'Lat', fill = 'Cantidad de meses') +
  ggtitle(label = glue('Cantidad de meses promedio por Estado y sexo'),
          subtitle = glue('Mujeres')) +
  theme_bw() +
  theme(legend.position = 'bottom',
        legend.key.width = unit(4, 'line'),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(family = 'Roboto', size = 50, face = 'bold',
                                  hjust = 0.5),
        plot.subtitle = element_text(family = 'Roboto', size = 50,
                                     face = 'bold', hjust = 0.5),
        axis.text.y = element_text(angle = 90, hjust = 0.5, family = 'Roboto',
                                   size = 30),
        axis.text.x = element_text(family = 'Roboto', size = 30),
        axis.title = element_text(family = 'Roboto', size = 40, face = 'bold'),
        legend.title = element_text(family = 'Roboto', face = 'bold',
                                    size = 50),
        legend.text = element_text(family = 'Roboto', size = 40))

ggsave(plot = g_muj_mnt, filename = './png/maps/average_pays/meses_mujeres.png',
       units = 'in', width = 9, height = 7, dpi = 300)
```

```
# Hombres ------------------------------------------------------------------\
  ##Cantidad de meses promedio por estado y sexo Hombres (carpeta average_pays)
g_hmb_mnt <- ggplot() +
  geom_sf(data = filter(mnts, sexo == 'Meses Hombres'), aes(fill = months),
          col = 'grey60', lwd = 0.5)  +
  scale_fill_gradientn(colors = brewer.pal(n = 9, name = 'YlOrBr')[3:9]) +
  # , limits = c(0, 261)
  geom_sf(data = mex1, fill = NA, col = 'grey50', lwd = 0.3) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +

  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD),
                  family = 'Roboto', size = 5) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  labs(x = 'Lon', y = 'Lat', fill = 'Cantidad de meses') +
  ggtitle(label = glue('Cantidad de meses promedio por Estado y sexo'),
          subtitle = glue('Hombres')) +
  theme_bw() +
  theme(legend.position = 'bottom',
        legend.key.width = unit(4, 'line'),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(family = 'Roboto', size = 50, face = 'bold',
                                  hjust = 0.5),
        plot.subtitle = element_text(family = 'Roboto', size = 50,
                                     face = 'bold', hjust = 0.5),
        axis.text.y = element_text(angle = 90, hjust = 0.5, family = 'Roboto',
                                   size = 30),
        axis.text.x = element_text(family = 'Roboto', size = 30),
        axis.title = element_text(family = 'Roboto', size = 40, face = 'bold'),
        legend.title = element_text(family = 'Roboto', face = 'bold',
                                    size = 50),
        legend.text = element_text(family = 'Roboto', size = 40))

ggsave(plot = g_hmb_mnt, filename = './png/maps/average_pays/meses_hombre.png',
       units = 'in', width = 9, height = 7, dpi = 300)
```

3. Figura 12: Elaboración propia, clasificación espacial según indice de LISA de los pagos de siniestros de diabetes.

```
# Hombres ----------------------------------------------------------------
clrs_lbls <- clrs$color
names(clrs_lbls) <- clrs$clase
crds <- hmbr %>% filter(!cluster %in% c('Not significant', 'Isolated')) %>%
  st_centroid() %>% st_coordinates() %>% as_tibble() %>% mutate(ENTIDAD =
                    pull(filter(hmbr, !cluster %in% c('Not significant',
                                                      'Isolated')), ENTIDAD))
g_hmb <- ggplot() +
  geom_sf(data = hmbr, aes(fill = clase), col = 'grey60', lwd = 0.5)  +
  scale_fill_manual(values = clrs_lbls, na.value = 'green') +
  geom_sf(data = mex1, fill = NA, col = 'grey50', lwd = 0.3) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +
  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD), family =
                    'Roboto', size = 5) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  labs(x = 'Lon', y = 'Lat', fill = 'Clúster') +
  ggtitle(label = glue('Clasificación espacial según índice LISA'),
          subtitle = glue('Hombres')) +
  theme_bw() +
  theme(legend.position = 'bottom',
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(family = 'Roboto', size = 50, face = 'bold',
                                  hjust = 0.5),
        plot.subtitle = element_text(family = 'Roboto', size = 50, face =
                                       'bold', hjust = 0.5),
        axis.text.y = element_text(angle = 90, hjust = 0.5, family = 'Roboto',
                                   size = 30),
        axis.text.x = element_text(family = 'Roboto', size = 30),
        axis.title = element_text(family = 'Roboto', size = 40, face = 'bold'),
        legend.title = element_text(family = 'Roboto', face = 'bold',
                                    size = 50),
        legend.text = element_text(family = 'Roboto', size = 40))
ggsave(plot =g_hmb,filename ='./png/maps/moran_average_pays/moran_hombres.png',
       units = 'in', width = 9, height = 7, dpi = 300)
```

```
# Mujeres ----------------------------------------------------------------
clrs_lbls <- clrs$color
names(clrs_lbls) <- clrs$clase
crds <- mjrs %>% filter(!cluster %in% c('Not significant',
                'Isolated')) %>% st_centroid() %>% st_coordinates()
                 %>% as_tibble() %>% mutate(ENTIDAD = pull(filter(mjrs,
                   !cluster %in% c('Not significant', 'Isolated')), ENTIDAD))
g_mjr <- ggplot() +
  geom_sf(data = mjrs, aes(fill = clase), col = 'grey60', lwd = 0.5)  +
  scale_fill_manual(values = clrs_lbls, na.value = 'green') +
  geom_sf(data = mex1, fill = NA, col = 'grey50', lwd = 0.3) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +
  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD), family =
                    'Roboto', size = 5) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  labs(x = 'Lon', y = 'Lat', fill = 'Clúster') +
  ggtitle(label = glue('Clasificación espacial según índice LISA'),
          subtitle = glue('Mujeres')) +
  theme_bw() +
  theme(legend.position = 'bottom',
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(family = 'Roboto', size = 50, face = 'bold',
                                  hjust = 0.5),
        plot.subtitle = element_text(family = 'Roboto', size = 50, face =
                                       'bold', hjust = 0.5),
        axis.text.y = element_text(angle = 90, hjust = 0.5, family = 'Roboto',
                                   size = 30),
        axis.text.x = element_text(family = 'Roboto', size = 30),
        axis.title = element_text(family = 'Roboto', size = 40, face = 'bold'),
        legend.title =element_text(family='Roboto', face='bold', size = 50),
        legend.text = element_text(family = 'Roboto', size = 40))
ggsave(plot  = g_mjr, filename =
         './png/maps/moran_average_pays/moran_mujeres.png',
       units = 'in', width = 9, height = 7, dpi = 300)
```

4. Figura 13: Elaboración propia, clasificación espacial según indice LISA de la duración media en meses.

```
# Meses ------------------------------------------------------------------
crds <- hmbr_mnts %>% filter(!cluster %in% c('Not significant', 'Isolated'))
%>% st_centroid() %>% st_coordinates() %>% as_tibble() %>% mutate(ENTIDAD =
                      pull(filter(hmbr_mnts, !cluster %in% c('Not significant',
                                                             'Isolated')), ENTIDAD))
g_hmb_mnt <- ggplot() +
  geom_sf(data = hmbr_mnts, aes(fill = clase), col = 'grey60', lwd = 0.5)  +
  scale_fill_manual(values = clrs_lbls, na.value = 'green') +
  geom_sf(data = mex1, fill = NA, col = 'grey50', lwd = 0.3) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +
  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD), family =
                    'Roboto', size = 5) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  labs(x = 'Lon', y = 'Lat', fill = 'Clúster') +
  ggtitle(label = glue('Clasificación espacial según índice LISA'),
          subtitle = glue('Cantidad de meses para los Hombres')) +
  theme_bw() +
  theme(legend.position = 'bottom',
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(family = 'Roboto', size = 50, face = 'bold',
                                  hjust = 0.5),
        plot.subtitle = element_text(family = 'Roboto', size = 50, face =
                                       'bold', hjust = 0.5),
        axis.text.y = element_text(angle = 90, hjust = 0.5, family = 'Roboto',
                                   size = 30),
        axis.text.x = element_text(family = 'Roboto', size = 30),
        axis.title = element_text(family = 'Roboto', size = 40, face = 'bold'),
        legend.title = element_text(family = 'Roboto', face = 'bold',
                                    size = 50),
        legend.text = element_text(family = 'Roboto', size = 40))
ggsave(plot=g_hmb_mnt, filename =
         './png/maps/moran_average_pays/moran_hombres_months.png',
       units = 'in', width = 9, height = 7, dpi = 300)
```

```
# Meses ------------------------------------------------------------------------
crds <- mjrs_mnts %>% filter(!cluster %in% c('Not significant', 'Isolated'))
%>% st_centroid() %>% st_coordinates() %>% as_tibble() %>% mutate(ENTIDAD
                    = pull(filter(mjrs_mnts, !cluster %in% c('Not significant',
                                                             'Isolated')), ENTIDAD))
mjrs_mnts
g_mjr_mnt <- ggplot() +
  geom_sf(data = mjrs_mnts, aes(fill = clase), col = 'grey60', lwd = 0.5)  +
  scale_fill_manual(values = clrs_lbls, na.value = 'green') +
  geom_sf(data = mex1, fill = NA, col = 'grey50', lwd = 0.3) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +
  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD), family =
                    'Roboto', size = 5) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  labs(x = 'Lon', y = 'Lat', fill = 'Clúster') +
  ggtitle(label = glue('Clasificación espacial según índice LISA'),
          subtitle = glue('Cantidad de meses para las Mujeres')) +
  theme_bw() +
  theme(legend.position = 'bottom',
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(family = 'Roboto', size = 50, face = 'bold',
                                  hjust = 0.5),
        plot.subtitle = element_text(family = 'Roboto', size = 50, face ='bold',
                                     hjust = 0.5),
        axis.text.y = element_text(angle = 90, hjust = 0.5, family = 'Roboto',
                                   size = 30),
        axis.text.x = element_text(family = 'Roboto', size = 30),
        axis.title = element_text(family = 'Roboto', size = 40, face = 'bold'),
        legend.title = element_text(family = 'Roboto', face = 'bold',
                                    size = 50),
        legend.text = element_text(family = 'Roboto', size = 40))
ggsave(plot = g_mjr_mnt, filename =
          './png/maps/moran_average_pays/moran_mujeres_months.png',
       units = 'in', width = 9, height = 7, dpi = 300)
'
```

5. Figura 15: Elaboración propia, mapa de frecuencia por estado de la republica mexicana.

```
# World shapefile ---------------------------------------------------------
wrld <- ne_countries(returnclass = 'sf', scale = 50)

# To make the map (conteo de siniestros por estado) -----------------------
g_conteo <- ggplot() +
  geom_sf(data = cntn, aes(fill = Freq)) +
  scale_fill_gradientn(colors = brewer.pal(n = 9, name = 'YlOrBr')) +
  geom_sf(data = shpf, fill = NA, col = 'grey40', lwd = 0.5) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  labs(x = 'Lon', y = 'Lat', fill = 'Cantidad siniestros') +
  theme_bw() +
  theme(legend.position = 'bottom',
        legend.key.width = unit(4, 'line'),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(family = 'Roboto', size = 50, face = 'bold',
                                  hjust = 0.5),
        plot.subtitle = element_text(family = 'Roboto', size =50, face ='bold',
                                     hjust = 0.5),
        axis.text.y = element_text(angle = 90, hjust = 0.5, family = 'Roboto',
                                   size = 30),
        axis.text.x = element_text(family = 'Roboto', size = 30),
        axis.title = element_text(family = 'Roboto', size = 40, face = 'bold'),
        legend.title = element_text(family = 'Roboto', face ='bold', size =50),
        legend.text = element_text(family = 'Roboto', size = 40))

ggsave(plot = g_conteo, file ='png/maps/general/conteo_siniestros_estados.png',
       units = 'in', width = 10.5, height = 7, dpi = 300)
```

6. Figura 17: Elaboración propia, contraste de los pagos de los hombres por estado con su p-valor por estado.

```
# Models ---------------------------------------------------------------------
mdls_hmbr <- mdls %>% filter(sexo == 'Masculino') %>% filter(estadistico ==
                                                               'duracion_m')
mdls_hmbr <- inner_join(shpf, mdls_hmbr, by = c('ENTIDAD' = 'estado'))
mdls_hmbr
plot(dplyr::select(mdls_hmbr, estimado))

crds <- mutate(as.data.frame(st_coordinates(st_centroid(mdls_hmbr))), ENTIDAD =
                 mdls_hmbr$ENTIDAD)

# To make the map
g_mdls_hmbr <- ggplot() +
  geom_sf(data = mdls_hmbr, aes(fill = estimado)) +
  scale_fill_gradientn(colors = brewer.pal(n = 9, name = 'Reds')) +
  geom_sf(data = shpf, fill = NA, col = 'grey40', lwd = 0.5) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +
  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD), family =
                    'Roboto', size = 2) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  ggtitle(label = 'Cantidad de pago estimada por cada mes de padecimiento') +
  labs(x = 'Lon', y = 'Lat', fill = 'Pago ($MXN)') +
  theme_bw() +
  theme(legend.position = 'bottom',
        legend.key.width = unit(4, 'line'),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        )
g_mdls_hmbr
dir_create('png/maps/linearModel')
ggsave(plot = g_mdls_hmbr, filename =
         glue('./png/maps/linearModel/pagos_hombres.png'), units = 'in',
       width = 9, height = 7, dpi = 300)
# Add the pvalue map
g_pvle_hmbr <- ggplot() +
  geom_sf(data = mdls_hmbr, aes(fill = p_value)) +
  scale_fill_gradientn(colors = rev(brewer.pal(n = 9, name = 'Greens'))) +
  geom_sf(data = shpf, fill = NA, col = 'grey40', lwd = 0.5) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +
  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD), family =
                    'Roboto', size = 2) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  ggtitle(label = 'Estadístico p-valor') +
  labs(x = 'Lon', y = 'Lat', fill = 'P-valor') +
  theme_bw() +
  theme(legend.position = 'bottom',
        legend.key.width = unit(4, 'line'),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        # plot.title = element_text(family = 'Roboto', size = 50, face ='bold',
        hjust = 0.5),
        )
g_pvle_hmbr
dir_create('png/maps/linearModel')
ggsave(plot = g_pvle_hmbr, filename =
         glue('./png/maps/linearModel/pagos_hombres_pvalue.png'), units = 'in',
       width = 9, height = 7, dpi = 300)

g_hmbr <- ggarrange(g_mdls_hmbr, g_pvle_hmbr, ncol = 2, nrow = 1)
ggsave(plot = g_hmbr, filename =
         glue('./png/maps/linearModel/Pago_hombres_pvalue_model.png'), units =
         'in', width = 14, height = 7, dpi = 300)
g_hmbr
```

7. Figura 18: Elaboración propia, contraste de los pagos de los mujeres por estado con su p-valor por estado.

```
# Models ------------------------------------------------------------------
mdls_mjr<-mdls %>%filter(sexo=='Femenino')%>%filter(estadistico =='duracion_m')
mdls_mjr <- inner_join(shpf, mdls_mjr, by = c('ENTIDAD' = 'estado'))
mdls_mjr
plot(dplyr::select(mdls_mjr, estimado))

crds <- mutate(as.data.frame(st_coordinates(st_centroid(mdls_mjr))), ENTIDAD =
                 mdls_mjr$ENTIDAD)
g_mdls_mjrs <- ggplot() +
  geom_sf(data = mdls_mjr, aes(fill = estimado)) +
  scale_fill_gradientn(colors = brewer.pal(n = 9, name = 'Reds')) +
  geom_sf(data = shpf, fill = NA, col = 'grey40', lwd = 0.5) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +
  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD), family =
                    'Roboto', size = 2) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  ggtitle(label = 'Cantidad de pago estimada por cada mes de padecimiento') +
  labs(x = 'Lon', y = 'Lat', fill = 'Pago ($MXN)') +
  theme_bw() +
  theme(legend.position = 'bottom',
        legend.key.width = unit(4, 'line'),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        )
g_mdls_mjrs
ggsave(plot = g_mdls_mjrs, filename =
         glue('./png/maps/linearModel/pagos_mujeres.png'), units = 'in',
       width = 9, height = 7, dpi = 300)
# Add the pvalue map
g_pvle_mjr <- ggplot() +
  geom_sf(data = mdls_mjr, aes(fill = p_value)) +
  scale_fill_gradientn(colors = rev(brewer.pal(n = 9, name = 'Greens'))) +
  geom_sf(data = shpf, fill = NA, col = 'grey40', lwd = 0.5) +
  geom_sf(data = wrld, fill = NA, col = 'grey40', lwd = 0.2) +
  geom_text_repel(data = crds, aes(x = X, y = Y, label = ENTIDAD), family =
                    'Roboto', size = 2) +
  coord_sf(xlim = ext(mex1)[1:2], ylim = ext(mex1)[3:4]) +
  ggtitle(label = 'Estadístico p-valor') +
  labs(x = 'Lon', y = 'Lat', fill = 'P-valor') +
  theme_bw() +
  theme(legend.position = 'bottom',
        legend.key.width = unit(4, 'line'),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        )
dir_create('png/maps/linearModel')
ggsave(plot = g_pvle_mjr, filename =
         glue('./png/maps/linearModel/pagos_mujeres_pvalue.png'), units = 'in',
       width = 9, height = 7, dpi = 300)
g_mjr <- ggarrange(g_mdls_mjrs, g_pvle_mjr, ncol = 2, nrow = 1)
g_mjr
ggsave(plot = g_mjr, filename =
         glue('./png/maps/linearModel/Pago_mujeres_pvalue_model.png'),
       units = 'in', width = 14, height = 7, dpi = 300)
```