UNIVERSIDAD DE SONORA
DIVISIÓN DE CIENCIAS EXACTAS Y NATURALES
MAESTRÍA EN CIENCIA DE DATOS

Proyecto Final de Probabilidad y Estadística: Formulación matemática del Cociente de Localización (*location quotient*) para la estimación de concentración industrial

Asignatura: Matemáticas para Ciencia de Datos

Profesora: Doctora Gudelia Figueroa Preciado

Alumna: María Elena Martínez Manzanares

Hermosillo, Sonora a 15 de noviembre del 2022

Contenido

Antecedentes

Dado un conjunto de objetos etiquetados en una cantidad finita de clases distribuidos en un espacio separado por bloques o secciones, los cocientes de localización, generalmente conocidos por su nombre en inglés, *location quotients*, son índices que permiten el análisis de concentración de objetos de una misma clase en un determinado bloque. Dependiendo de la naturaleza de los objetos, clases y bloques espaciales, diferentes índices pueden ser propuestos para análisis (c.f. Ellison y Gleaser (1997)).

En el caso de cocientes de localización para medir concentración industrial, LQ, la U.S. Bureau of Labor Statistics utiliza el cociente

$$LQ = \frac{\frac{total\ negocios\ de\ giro\ k\ en\ una\ región}{total\ negocios\ en\ la\ región}}{\frac{total\ negocios\ de\ giro\ k\ en\ el\ espacio\ universo}{total\ negocios\ en\ el\ espacio\ universo}}. \quad (1)$$

Billings y Johnson (2012) mencionan que, para la época, pocos intentos se habían efectuado para derivar (1) a partir de un modelo matemático formal que permitiese cuantificar por medio de herramientas estadísticas la confiabilidad del cociente para reflejar la densidad industrial por zonas. Motivados por esto, Billings y Johnson (2012) formulan el modelo matemático que permite el análisis de (1) por medio de procesos binomiales y de Poisson.

Objetivos

- Explicar la fundamentación teórica planteada en Billings y Johnson (2012).
- Calcular a nivel estado el cociente de localización de los giros industriales reconocidos por el Sistema de Clasificación Industrial de América del Norte (o por sus siglas en inglés, NAICS).
- Replicar el análisis de sesgo del cociente de localización.
- Analizar la confiabilidad del cociente de localización calculado a nivel estado.

Descripción de variables

Definiremos como K y J la cantidad de giros industriales reconocidos por el NAICS y la cantidad de bloques que segmentan un espacio que consideremos como universo, respectivamente. Por ejemplo, puede ser considerado el universo como el país completo de Estados Unidos y los bloques que segmentan los estados o ciudades, o considerar como universo el estado y los bloques que segmentan las ciudades que componen al estado.

Definimos como s_{jk} la variable aleatoria (v.a.) que representa la cantidad de establecimientos industriales en el bloque $j \in \{1,\dots,J\}$ del giro industrial $k \in \{1,\dots,K\}$. La v.a. s_k será la cantidad de establecimientos industriales del giro k en el universo.

Por otro lado, definimos a x_j la variable aleatoria que representa la cantidad total de establecimientos industriales en el bloque j y x la v.a. que denota la cantidad de establecimientos industriales en el universo. Por lo tanto, $\sum_{j=1}^{J} x_j = x$.

Finalmente, definimos la distribución empírica sobre la cantidad de establecimientos industriales en el universo como

$$\theta_j := \frac{x_j}{x}, j \in \{1,\dots,J\}.$$

Descripción de fuente de datos

Los análisis efectuados en este trabajo fueron realizados a partir de la base de datos County Business Patterns del año 2000 el cual contiene la cantidad industrias clasificadas por medio del NAICS agrupadas por ciudad, específicamente se hizo uso de la tabla llamada est_cnty2000 extraído de Holmes (2003a). El diccionario de la tabla est_cnty2000 puede ser visto en la Tabla 1.

Como se menciona en la Tabla 1, el NAICS se compone de seis dígitos que identifican diferentes giros industriales. Al respecto, el U.S. Bureau of Economic Analysis (2005) menciona que los primeros dos dígitos representan el sector, el tercero el subsector, el cuarto el grupo industrial, el quinto la industria NAICS y el sexto reconoce la industria nacional. Concretamente, el California State University Chico Meriam Library (2016) presenta el siguiente ejemplo

Tabla 1: Descripción de las características de la tabla est_cnty2000.

Nombre de la variable	Descripción
NAICS	6 dígitos del NAICS
ST	Número del estado
COUNTY	Número de la ciudad
EMPCAT	Clasificación de cantidad de estados. Específicamente ‘01’=1-4 ‘02’=5-9 ‘03’=10-19 ‘04’=20-49 ‘05’=50-99 ‘06’=100-249 ‘07’=250-499 ‘08’=500-999 ‘09’=1,000-1,499 ‘10’=1,500-2,499 ‘11’=2,500-4,999 ‘12’=5,000+
NUMBER	Frecuencia de establecimientos en el registro

Fuente: Elaboración propia a partir de Holmes (2003b).

44 – Retail
445 - Food & Beverage Stores
4452 - Specialty Food Stores
445291 - Baked Goods Stores.

Por lo anterior, truncamientos de los seis dígitos nos presentan des agrupaciones dentro de un mismo giro industrial.

Cociente de localización

Planteamiento teórico

Se supondrá que la decisión sobre locaciones de establecimientos comerciales y la actual densidad de locaciones comerciales son independientes.

Dado el giro industrial k y la zona de análisis j, supondremos que la cantidad de empresas de giro industrial k en la j-ésima zona de análisis, s_{kj}, es una variable aleatoria que sigue una distribución binomial con parámetros (s_k, θ_j) y rango $\{0,1,2,\dots,s_k\}$. Diremos que la terna

$$(s_{kj}, s_k, \theta_j)_{j,k} \tag{2}$$

definen el proceso binomial. Notemos además que $E[s_{jk}] = s_k\theta_j =: \lambda_{jk}$. Definimos el cociente de localización del giro industrial k en la j-ésima zona de análisis como

$$LQ_{jk} := \frac{s_{jk}}{\lambda_{jk}}. \tag{3}$$

Se sigue inmediatamente que $E[LQ_{jk}] = 1$, es decir, es un estimador insesgado independientemente del tamaño de la zona de análisis y de la cantidad de empresas del giro industrial. Por otro lado, podemos notar que el cociente es un índice monótono, e interpretable en el sentido de que valores mayores a uno en (3) demuestran que se contabilizaron una cantidad mayor de negocios en la zona y del giro que la cantidad esperada para esa zona en ese giro.

Es un resultado conocido que, en el caso de analizar sistemas de grandes poblaciones, el proceso binomial se aproxima a un proceso de Poisson. Concretamente, simplificando la notación a $s_{jk} = s$, $s_k = S$, y $\theta_j = \theta$, tenemos que si $S \to \infty$, el proceso binomial (2) se convierte en un proceso de Poisson con parámetro $\lambda := \lim_{S\to\infty} S\theta$. Una importante propiedad del cociente de localización visto como un proceso Poisson es que es insesgado y conserva monotonía (ver Proposition 2.2, Billings y Johnson (2012)).

En la práctica, cuando se considera que el tamaño de S es suficientemente grande, la estimación de λ se realiza a través de $\bar{\lambda} = S\theta$, lo cual ocasiona que el estimador (2) sea sesgado. En efecto, notemos que la esperanza de LQ puede descomponerse como la suma de los primeros S elementos, $\{0,1,2,\dots,S\}$ y la suma partiendo de $S+1$, $\{S+1,\dots\}$, de la siguiente manera

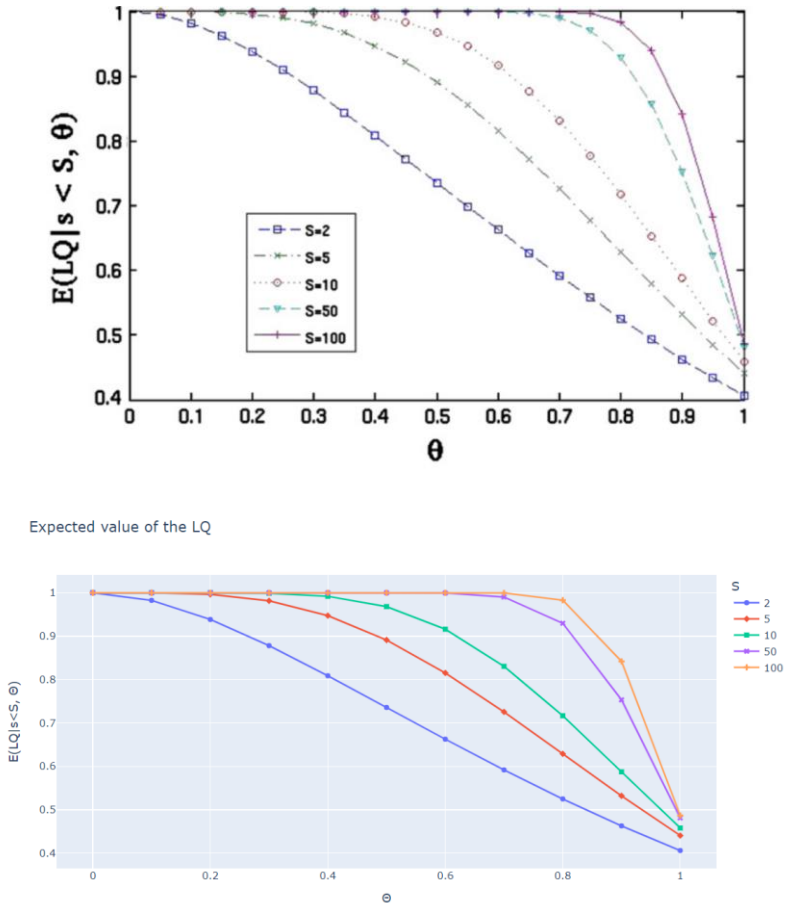
$$\begin{aligned} E[LQ]_0^\infty &= E[LQ]_0^S + E[LQ]_{S+1}^\infty, \\ &= e^{-\lambda}\left(\sum_{k=0}^{S-1}\frac{\lambda^k}{k!} + \sum_{k=S+1}^{\infty}\frac{\lambda^k}{k!}\right). \end{aligned} \tag{4}$$

Considerar un $\bar{\lambda} = S\theta$ implica modelar un sistema de Poisson con rango truncado al conjunto $\{0,1,2,\ldots,S\}$. Si tenemos S suficientemente grande, se tendrá que $E[LQ]_0^S \approx E[LQ]_0^\infty$, pero $E[LQ]_0^S < 1$.

Billings y Johnson (2012) presentan una gráfica donde cuantifican la magnitud del sesgo calculando $E[LQ]_0^S$ con diferentes parámetros de S y θ. Para efectos de este trabajo, esta gráfica fue replicada en Python por medio de la librería de SciPy. Los resultados se muestran en la Figura 1.

Una vez establecido lo anterior, Billings y Johnson (2012) plantea la situación de determinar puntos críticos para la estimación de intervalos de confianza de la variable aleatoria LQ. Un valor critico $\bar{s}$ de la distribución $f(s|s,\theta)$ dado un nivel de significancia α

Figura 1: Estimación del sesgo del cociente de localización considerando diferentes valores de S y θ. Arriba la figura presentada en Billings y Johnson (2012), abajo la replicación de la figura realizada en Python.



Fuente: Elaboración propia a partir de Billings y Johnson (2012).

de una cola se puede determinar por medio de la solución de

$$\min_{\bar{s}} \bar{s} \ \ sujeto\ a \ \sum_{s=0}^{\bar{s}} f(s|S,\theta) \geq (1-\alpha). \tag{5}$$

Debido a que el proceso que define (5) es discreto, es un hecho conocido que la solución (5) admite el uso de diferentes niveles de significancia definidos en un intervalo el cual contiene el nivel de significancia original dado α. La amplitud de este intervalo queda determinada por la magnitud de $P[s = \bar{s}|S,\theta]$. Definimos $\tau := P[s = \bar{s}|S,\theta]$ como el nivel de tolerancia, el cual se puede interpretar como la amplitud de este intervalo.

En la práctica, el nivel de tolerancia solamente puede ser estimado; sea $\hat{\tau}$ esta estimación de τ. Billings y Johnson (2012) mencionan que en el caso de suponer que el proceso es de Poisson, puede determinarse el valor mínimo del parámetro $\bar{\lambda}$ que permita tener un nivel de significancia α con una tolerancia $\hat{\tau}$ por medio de un grid search sobre los valores de $\bar{\lambda}$. En la Tabla 1 se contrastan los resultados obtenidos por un grid search con valores $\bar{\lambda}$ en el rango $\{1,2,3,\dots,400\}$ con el grid search presentado por Billings y Johnson (2012). La relevancia de los datos presentados en la Tabla 2 es que nos permiten determinar los valores mínimos de $\bar{\lambda}$ que permiten tener un buen estimador del cociente de localización LQ.

Replica de resultados a nivel estado

Billings y Johnson (2012) presentan el porcentaje de giros industriales reconocidos por el NAICS que satisfacen tener $\bar{\lambda}$ mayor o igual a los presentados en la Tabla 2 (superior) con niveles de agregación a nivel código postal, ciudad y estado y considerando dos, cuatro y seis dígitos del NAICS.

Tabla 2: Las celdas contienen valores de $\bar{\lambda}$ que garantizan niveles de tolerancia $\hat{\tau}$ y nivel de confianza $(1-\alpha)$. En la parte de arriba se presentan los valores expuestos en Billings y Johnson (2012), en la parte de abajo son los resultados obtenidos con un grid search de rango $\{1,2,3,\dots,400\}$ sobre $\bar{\lambda}$.

Tolerancia ($\hat{\tau}$)	Nivel de confianza $(1-\alpha)$		
	0.9	0.95	0.99
0.01	315	112	2.9
0.025	51	19	1.8
0.05	13.7	4.7	0.44

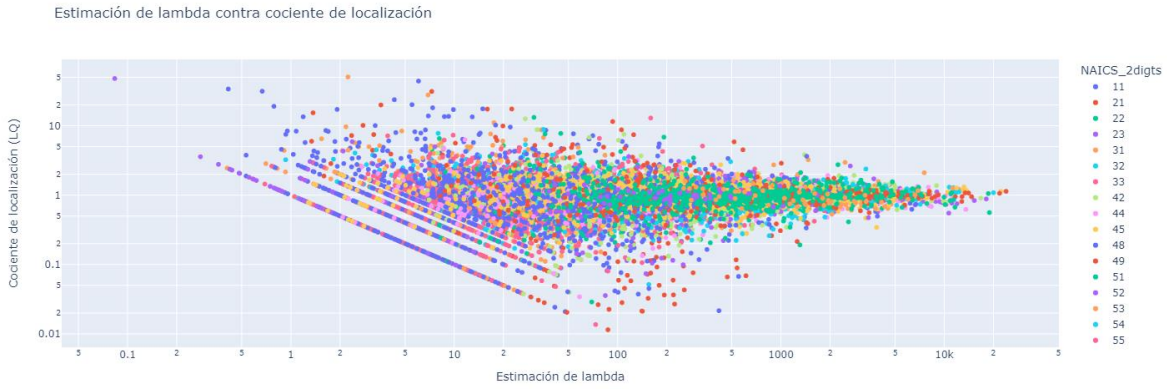
Tolerancia ($\hat{\tau}$)	Nivel de confianza $(1-\alpha)$		
	0.9	0.95	0.99
0.01	279	88	9
0.025	43	11	1
0.05	8	2	0

Fuente: Elaboración propia a partir de Billings y Johnson (2012).

Para el presente trabajo se realizó una réplica de los porcentajes obtenidos a nivel estado considerando los umbrales de $\bar{\lambda}$ presentados en la Tabla 2 (inferior) considerando truncamientos de cuatro dígitos del NAICS. Los resultados de este análisis se presentan en la Tabla 3.

Finalmente, se calcularon los cocientes de localización a partir de (3) donde se obtuvieron los resultados de la Figura 2.

Figura 2: Se presentan diferentes variables que determinan el cociente de localización. Las relaciones lineales que se aprecian son derivadas del cociente que define al cociente de localización. Se realizo un matizado a partir de 2 dígitos del NAICS pero los cálculos fueron realizados a partir de 4 dígitos. La gráfica 1 y 2, y 3 y 4 presentan la misma información con la dimensión de la cantidad de industrias que hay en ese giro a nivel país incluida.



Fuente: Elaboración propia.

Interpretación de resultados

La información porcentual calculada en la Tabla 3 nos permite determinar que, a nivel estado, el cociente de localización definido como en (3) resulta en un estimador con bajo nivel de sesgo para agregaciones de 4 dígitos del NAICS. Del análisis de los cocientes de localización calculados no se detectaron relaciones que no de derivaran directamente de la definición (3), salvo una presencia importante de empresas con el NAICS de dos dígitos 81, la cual se corresponde con el giro "otro servicio", lo cual podría explicar su frecuencia de aparición en los registros.

Conclusiones

En el presente trabajo se expusieron las formulaciones teóricas planteadas por Billings y Johnson (2012) que dan el sustento teórico necesario para analizar la confiabilidad del cociente de localización utilizando por el U.S. Bureau of Labor Statistics para determinar la densidad industrial en Estados Unidos en diferentes zonas del país considerando los diferentes giros industriales reconocidos por el NAICS. Se puede concluir de Billings y Johnson (2012) que el cociente de localización presenta ser un estimador poco confiable para niveles de segmentación a nivel código postal y con truncamientos de 4 y 6 dígitos de NAICS si se considera que los datos son generados a través de un proceso de Poisson.

Es debido a este análisis que se conjetura que para estudios de densidad industrial con altos niveles de segmentación, como lo son a nivel código postal, y/o con segmentación de los giros industriales a 6 dígitos del NAICS, realizar un cálculo del cociente de localización suponiendo que los datos son generados a través de un proceso Binomial puede mejorar la confiabilidad del cálculo del cociente de localización.

Tabla 3: Las celdas contienen el porcentaje de empresas con agregación a nivel estado con giros industriales de cuatro dígitos del NAICS que tienen un valor mayor o igual de $\bar{\lambda}$ para los niveles de tolerancia y de confianza. En la parte de arriba se presentan los valores expuestos en Billings y Johnson (2012), en la parte de abajo son los resultados obtenidos en la replicación.

Tolerancia ($\hat{\tau}$)		
0.01	0.025	0.05
Nivel de confianza $(\mathbf{1-\alpha=0.90})$		
69.58%	94.48%	99.27%
Nivel de confianza $(\mathbf{1-\alpha=0.95})$		
87.64%	98.98%	99.90%
Nivel de confianza $(\mathbf{1-\alpha=0.99})$		
99.76%	99.98%	100.00%

Tolerancia ($\hat{\tau}$)		
0.01	0.025	0.05
Nivel de confianza $(\mathbf{1-\alpha=0.90})$		
89.61%	98.71%	99.89%
Nivel de confianza $(\mathbf{1-\alpha=0.95})$		
97.08%	99.82%	99.98%
Nivel de confianza $(\mathbf{1-\alpha=0.99})$		
99.98%	99.99%	100.00%

Fuente: Elaboración propia a partir de Billings y Johnson (2012).

Anexo

Código fuente de los análisis

Los análisis efectuados fueron realizados en Python a través de la plataforma Colab de Google, y se dejan a disposición del lector para su revisión a través de la liga https://bit.ly/Manzanares_AnalisisLQ o escaneando el siguiente código QR.

Referencias

I. Billings, S. B., & Johnson, E. B. (2012). The location quotient as an estimator of industrial concentration. *Regional Science and Urban Economics*, 42(4), 642-647.

II. Ellison, G., & Glaeser, E. L. (1997). Geographic concentration in US manufacturing industries: a dartboard approach. *Journal of political economy*, *105*(5), 889-927.

III. Holmes, T. (2003a). *County Business Patterns Establishment Data Page*. Universidad de Minnesota. http://users.econ.umn.edu/~holmes/data/CBP/

IV. Holmes, T. (2003b). *Documentation for County Business Patterns Establishment Data Page*. Universidad de Minnesota. https://users.econ.umn.edu/~holmes/data/CBP/Documentation_est_CBP.htm

V. State University Chico Meriam Library. (2016). *NAICS (North American Industry Classification System). What are NAICS codes?*. https://libguides.csuchico.edu/c.php?g=414121&p=2821979.

VI. U.S. Bureau of Economic Analysis. (16 de noviembre del 2005). *What is the difference between 2, 3, 4, 5, and 6-digit NAICS codes?*. https://www.bea.gov/help/faq/19.