

V. Aplicación de metodología estadística para la clasificación económica de la ciudad de Xalapa

JUAN ANDRÉS SÁNCHEZ GARCÍA²

Resumen

Las ciudades contemporáneas reflejan una serie de interacciones y fenómenos que necesitan ser estudiados para comprender la dinámica que ocurre en ellas. Una de las variables que marca una transformación en las ciudades es la económica, que convierte este sector en un punto de atención para explicar fenómenos de desarrollo en las urbes.

La actividad económica se convierte entonces en un epicentro de estudio cuyas técnicas no deben marcar solo un conteo de proporción en diferentes áreas, sino que se debe usar una metodología que explore y explique la clasificación económica con la menor cantidad de observaciones. Por tal motivo, este trabajo busca expresar la clasificación de las ciudades con una expresión estadística y, con base en la ciencia de datos, validar los residuos del modelo, así como explorar situaciones que pueden robustecer o fragilizar la ecuación que explique las actividades económicas significativas en el estudio.

Palabras clave: *clasificación de económica, ciudad, metodología estadística.*

² Doctor en Arquitectura, Diseño y Urbanismo. Profesor de tiempo completo de la Facultad de Arquitectura-Xalapa de la Universidad Veracruzana. ORCID: <http://orcid.org/0000-0003-2217-2711>; correo electrónico: juansanchez@uv.mx

Introducción

Las ciudades contemporáneas reflejan una naturaleza dinámica y multifacética, por lo que requieren ser estudiadas de manera interdisciplinaria a fin de comprender su complejidad, con base en la interacción de diversos componentes como la infraestructura, población, sociedad, economía, cultura, entre otros, por lo que su característica es precisamente la constante transformación y adaptación a los cambios, lo que las ha vuelto hasta cierto punto impredecibles.

Las ciudades eventualmente muestran cambios en las composiciones espaciales según las perspectivas de análisis para averiguar la introducción a la complejidad urbana lo que permite adentrarse en la comprensión de los fenómenos inherentes a las ciudades, sus interacciones y su comportamiento dinámico. Este estudio aborda los elementos que hacen que las ciudades sean sistemas complejos, desde la multiplicidad de actores hasta la diversidad de variables que influyen en su funcionamiento. Se busca comprender cómo emergen patrones y fenómenos no lineales a partir de la interacción de múltiples componentes en un entorno urbano. En definitiva, la introducción nos proporciona el punto de partida para analizar la complejidad que caracteriza a las ciudades y los desafíos que ello implica bajo el análisis de ciertos fenómenos.

Uno de ellos es la población sistémica, que en el estudio de las ciudades se refiere al enfoque analítico que considera a los habitantes urbanos como parte de un sistema interconectado en el que las dinámicas sociales, económicas, culturales y ambientales interactúan de manera compleja; en otras palabras, la forma espacial en que interactúan la organización y composición de la población. Este concepto busca comprender cómo los fenómenos demográficos, la movilidad, las interacciones sociales y otros aspectos influyen y son influenciados por la estructura y el funcionamiento de las ciudades, por lo que se vuelve necesario explorar la perspectiva espacial para interactuar con los desafíos urbanos contemporáneos, y diseñar políticas públicas y estrategias de planificación que promuevan la sostenibilidad, la equidad y la calidad de vida.

Uno de los fenómenos recurrentes en el aspecto urbano, que abona a

la dinámica social, es el que está ligado a la economía, por lo que el uso de herramientas enfocados en la economía urbana detona estudios que se centran en las relaciones espaciales de población y las actividades económicas para el funcionamiento y el desarrollo de ciudades.

Función y clasificación las ciudades; un patrón económico

La economía de cada ciudad engloba diferentes sectores. Si bien cada uno tiene sus especificidades, el hecho de que sus procesos de producción abarquen un área geográfica limitada (el área urbana) conlleva intercambios. Asimismo, algunos sectores ya están relacionados entre sí desde el punto de vista técnico (por ejemplo, los servicios públicos, la infraestructura, la construcción y los residuos) (OIT, 2024).

La importancia de la actividad económica en las ciudades representa un reto para los Objetivos de Desarrollo Sostenible, cuyo objetivo 11, ciudades y comunidades sostenibles, recalca el “apoyar los vínculos económicos, sociales y ambientales positivos entre las zonas urbanas, periurbanas y rurales fortaleciendo la planificación del desarrollo nacional y regional” (NU, 2024, párr. 15), por lo que es importante potenciar la economía de las ciudades.

Para alcanzar dicho objetivo e impulsar la economía en las ciudades, se necesita determinar su clasificación y sobre cuál actividad económica están funcionando para mantener la estructura de las políticas públicas en el desarrollo territorial. Las ciudades presentan características que permiten identificar las diferencias entre ciudades; una de ellas es la *función*, es decir, la manera en que la población ejerce sus fuentes de trabajo predominantes, las cuales cambian de una ciudad a otra según su historia o zona geográfica, pero cabe mencionar que en su conjunto pueden presentar la característica de una región. La función de una ciudad también representa una forma de clasificarla y con ello entender la estructura y forma en que se compone (Sánchez García, 2021).

La *función* es un término tomado de la fisiología que compara a la ciudad con un organismo vivo; entonces, la *función urbana* se entiende como “cualquier actividad que los ciudadanos ejercen en beneficio de los habi-

tantes de la aglomeración principal y de aquellos que residen en su región de influencia” (Zárate Martín, 2012, p. 31); este principio, en opinión de los geógrafos Chabot y Beaujeu Garnier (1972), justifica la existencia de la ciudad y su desarrollo.

Algunos estudios para determinar la clasificación de las ciudades muestran la ideada por C. D. Harris en 1943, quien “a partir de la proporción de la población empleada en ciertas profesiones daba una interpretación cuantitativa” (Johnson, 1974, p. 105) y clasificó ciertas etiquetas planteadas por él; lo cuestionable en este planteamiento es que no siempre la mayor cantidad de población acumulada que trabaja en cierta actividad económica es la más significativa para determinar los desarrollos económicos, ya que entran otros criterios como el valor adquisitivo de cada actividad o importancia para la región.

Otros estudios que intentaron definir la clasificación de la función en ciudades como Nueva Zelanda fue el propuesto por L. Pownall en 1956, en *The Origins of Towns in New Zealand*, quien, a pesar de reconocer al empleo como un factor importante, concluyó que las ciudades de ese país eran meramente multifuncionales dada su homogeneidad, lo que abre una brecha grande para señalar que las estandarizaciones de homogeneidad son un factor clave para determinar las composiciones económicas en las regiones.

Ante esta propuesta, se concluye que los análisis, hasta cierto punto cuantitativos, son un rasgo importante en los estudios económicos. En trabajos como el de H. J. Nelson en Estados Unidos y W. Steigenga en Países Bajos se planteó el uso de la media aritmética como base de la clasificación; Nelson agrupó las profesiones de Estados Unidos calculando el porcentaje medio en cada categoría con base en el censo de 1950 y su desviación tipo correspondiente para cada grupo de profesionales, y es aquí donde se comienza a aportar criterios objetivos para la determinación de patrones de función en la ciudad mediante una estadística significativa, pero su trabajo no terminó aquí, sino que alcanzó a determinar la relación entre las características de la población y la especialización profesional (Johnson, 1974).

En este sentido, el estudio de la clasificación de la ciudad, con base en la actividad económica, se fundamenta en construcciones de evaluación y valoración de las urbes, acentuada por elementos medibles con ayuda de la interpretación de datos basados en inferencias con cálculos de probabilidades me-

dianthe la *estadística*. Lo importante de ser auxiliados por la estadística es saber discernir que el dato que se maneje o se obtenga durante la investigación carece de relevancia en el urbanismo si su interpretación no es la adecuada, es decir que, al trabajar con modelos e inferencias, deben ser estos solo el sustento de argumentos para explicar los fenómenos desde el urbanismo.

Al explorar la estadística como disciplina auxiliar es contundente afirmar que no es suficiente la gráfica, el modelo y la validación de supuestos, sino que es necesario visualizar la gama de procedimientos que de ella emana, por lo que se debiera explicar mejor en términos de *la ciencia de datos* como un campo multidisciplinario e interdisciplinario que contiene métodos científicos, procedimientos, técnicas y procesos que ayudan a entender y explicar los datos. La ciencia de datos ya no solo abarca a la estadística, sino que indaga en otras áreas como la minería de datos, la analítica predictiva o el aprendizaje automático. La *ciencia de datos* también puede entenderse como “un concepto para unificar estadísticas, análisis de datos, aprendizaje automático, y sus métodos relacionados, a efectos de comprender y analizar fenómenos reales” (Hayashi, 1998), lo que abarca más allá de la evaluación estadística.

Aunque se entiende que la ciencia de datos ha sido para muchos una disciplina de reciente creación, lo cierto es que John W. Tukey, en su artículo *The Future of Data Analysis* (1962), introdujo este término para explicar la evolución de la estadística matemática y definió al *análisis de datos* como los “procedimientos para analizar datos, técnicas para interpretar los resultados de dichos procedimientos, formas de planificar la recopilación de datos para hacer un análisis más fácil, más preciso o acertado”. Posteriormente, expone la necesidad del análisis de datos para sugerir hipótesis y generar modelos estadísticos en *Exploratory Data Analysis* (1977).

William S. Cleveland (2001), de la Purdue University, escribe el artículo “Data science: an action plan for expanding the technical areas of the field of statistics” donde explica que la ciencia de datos es una disciplina independiente que vincula a la estadística con las ciencias de la computación para trabajar en distintas investigaciones multidisciplinarias por medio de modelos y métodos de datos, computación de datos, enseñanza con los datos o incluso la generación de teorías con estos (Cleveland, 2001).

Lo importante es que la ciencia de datos engloba la necesidad de gene-

rar métodos como formas de análisis y modelos como formas de síntesis para ayudar a estudiar la realidad y los fenómenos que se incluyen, por lo que la certeza que promueve esta disciplina ayuda a cuantificar y medir las dinámicas que se producen en el territorio.

Proceso estadístico de función económica de Xalapa

Para establecer el parámetro de clasificación económica, se presenta la tabla 1 con las abreviaciones por utilizar en el análisis estadístico que son las actividades registradas en el Instituto Nacional de Estadística y Geografía (INEGI).

Para determinar la función económica de la ciudad de Xalapa con respecto a la población ocupada, se realiza un ejercicio taxonómico basado en la clasificación de Harris (*A Functional Classification of Cities in the United States*, 1943) (Aguilar, Gratzbord y Sánchez Crispín, 1996) y se retoma la definición estadística propuesta por Nelson (*A Service Classification of American Cities*, 1995) para determinar su función económica.

Xalapa presenta características económicas puntuales y se busca que la variable principal de actividad económica sea la que determine la función de la ciudad. Esta “es el conjunto de acciones realizadas por una unidad económica con el propósito de producir o proporcionar bienes y servicios que se intercambian por dinero u otros bienes o servicios” (INEGI, 2020), mientras que la unidad económica es el lugar o entidad donde se realizan las actividades económicas.

En un análisis exploratorio de los censos económicos 2014 y 2019 se determina, mediante porcentajes que los comportamientos de actividades económicas por sectores tienen un comportamiento muy similar.

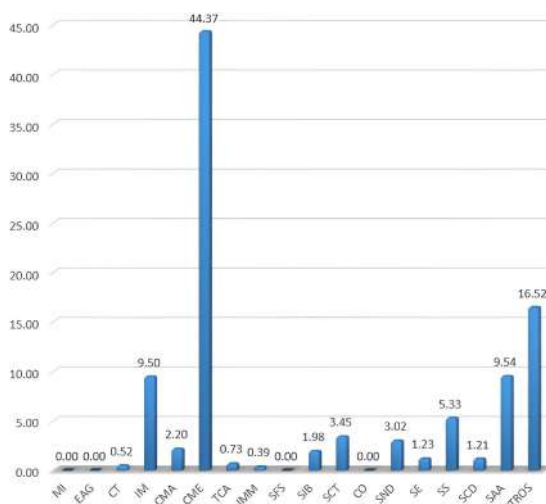
Tener la mayor cantidad de porcentaje de actividades económicas no quiere decir que la clasificación muestre una apariencia significativa, ya que no depende de la cantidad, sino del vínculo que se tenga con la cantidad de personas que trabajan en ellas; por ejemplo, hay actividades que aglomeran a más de 10 personas, mientras que otras solo tienen 1 o 2, por lo que el cruce de estas variables muestra la dependencia del sector productivo con respecto a las personas ocupadas en dichos trabajos.

Tabla 1 Abreviaciones para análisis

Variable	Abreviación
Unidad económica	UE
Población ocupada	PO
Actividad económica	AE
Minería	MI
Servicios financieros y de seguros	SFS
Corporativos	CO
Electricidad, agua y suministro de gas por ductos al consumidor final	EAG
Construcción	CT
Industrias manufactureras	IM
Comercio al por mayor	CMA
Comercio al por menor	CME
Transportes, correos y almacenamiento	TCA
Información en medios masivos	IMM
Servicios inmobiliarios y de alquiler de bienes muebles e intangibles	SIB
Servicios profesionales, científicos y técnicos	SCT
Servicios de apoyo a los negocios y manejo de desechos y servicios de remediación	SND
Servicios educativos	SE
Servicios de salud y de asistencia social	SS
Servicios de esparcimiento culturales y deportivos y otros servicios recreativos	SCD
Servicios de alojamiento temporal y de preparación de alimentos y bebidas	SAA

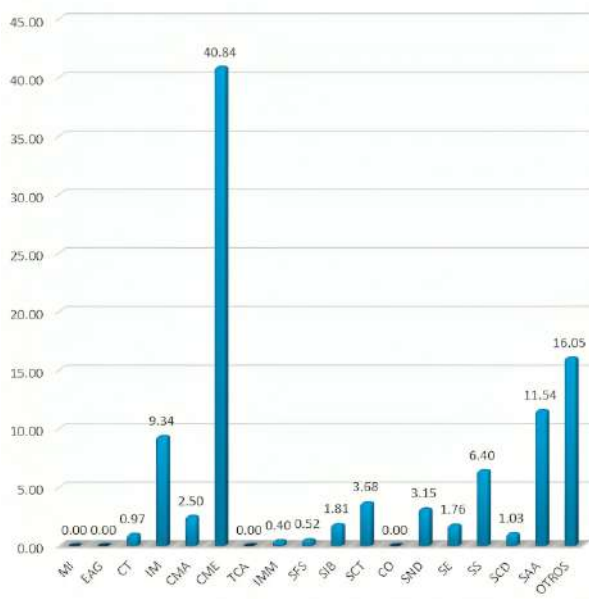
Fuente: elaboración propia.

Figura 1. Gráfica de barras del porcentaje de actividades económicas en el censo 2014



Fuente: elaboración propia con base en el censo económico 2014.

Figura 2. Gráfica de barras del porcentaje de actividades económicas en el censo 2009



Fuente: elaboración propia con base en el censo económico 2019 INEGI.

Para hacer el cruce entre unidades económicas y población ocupada, se busca una expresión matemática cuyo modelo de correlación es confiable para determinar al final la clasificación económica de la ciudad de Xalapa. Para ello, se utiliza una metodología de análisis multivariado referido al análisis de p variables respuesta medidas conjuntamente sobre n individuos. Generalmente, estas variables están asociadas y el análisis multivariado considera la estructura de asociación como un elemento importante que hace que los resultados obtenidos sean más adecuados a la realidad que aquellos que se obtendrían por el análisis univariado de cada una de las variables por separado (Ojeda, 2000).

Esta metodología permite analizar y clasificar dependencias y agrupación de objetos. Por tal motivo, el fenómeno de estudio comprende la correlación entre las actividades económicas (variables independientes X) con respecto a la población ocupada (variable dependiente Y).

Para ello se establece que, en una ciudad como Xalapa, las personas que pertenecen al sector de población ocupada están relacionadas con un

sector económico determinado por las diferentes actividades económicas, y esto puede ser determinado a través de métodos de inferencia. La inferencia estadística parte de la postulación, explícita o implícita, de un modelo que describe el comportamiento de una o varias variables aleatorias; este modelo está compuesto por dos partes: 1) parte sistemática (explicada); 2) parte aleatoria (no explicada) (Ojeda, 2000).

La representación de la ecuación estadística que se une para encontrar un modelo de dependencia se expresa de la siguiente manera:

Ecuación 1. *Modelo estadístico*

$$Y = f(A, B, \dots) + e(F, G)$$

A, B es un conjunto de factores o variables que se controlan, f es una función de los factores o variables, y e es una parte que describe la contribución de aleatoriedad que siempre se encuentra en los fenómenos estudiados debido a los factores $F, G, \dots$, que no se pueden controlar y a veces hasta son desconocidos.

Con estas condiciones de la ecuación se utiliza la técnica estadística de regresión lineal que se utiliza para modelar la relación entre variables. A fin de establecer causalidad, la relación entre las variables regresoras y las variables respuesta está sujeta a la lógica de la investigación y, partiendo del supuesto de tener 13 variables independientes (cada una de las UE) para explicar una sola variable dependiente (Y), se opta por utilizar una variante de este modelo que es la regresión lineal múltiple.

El modelo de regresión lineal múltiple es una extensión natural de la regresión lineal simple, al caso en el que se tiene más de una variable explicaría: es decir, el modelo es postulado considerando que a la respuesta Y contribuyen p variables explicativas $X_1, X_2, \dots, X_p$, y se expresa de la siguiente manera:

Ecuación 1. *Modelo de regresión lineal múltiple*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + e_i; i = 1, 2, \dots, n$$

lar el componente aleatorio e de este modelo estadístico se asume una variable aleatoria distribuida normalmente con media cero y varianza σ^2 ; es decir; $e \sim N(0, \sigma^2)$. Este supuesto también vincula que los errores son independientes con la misma distribución.

La hipótesis general con base la cual se construye el modelo es que las X contribuyen significativamente de manera conjunta para explicar la Y . Esta hipótesis de regresión implica el rechazo de la hipótesis nula, en favor de la alterna postulada (Ojeda, 2000), quedando de la siguiente manera:

$$H_o: \beta_j = 0 \text{ para } j = 1, 2, 3, \dots, p \text{ (} X_j \text{ no influye)}$$

vs

$$H_a: \beta_j \neq 0 \text{ para al menos una } j \text{ (} X_j \text{ si influye)}$$

Esta hipótesis se evalúa con la tabla de análisis de varianza valorando la R^2 y el p -value asociado a la F , así como el análisis de supuestos de homocedasticidad, independencia y normalidad en los residuos. Al correr el análisis de regresión en el programa MINITAB 17, se obtiene la siguiente ecuación ajustada de acuerdo con el método de mínimos cuadrados que minimiza la suma de los residuos al cuadrado:

*Ecuación 2. Modelo de regresión lineal múltiple entre la PO y las UE
(Sánchez, Argüello y Montano, 2018).*

$$\begin{aligned} PO = & 0.351 + 0.0042 \text{ SCD} - 0.0193 \text{ TCA} + 0.0070 \text{ SIB} - 0.0353 \text{ IMM} + 0.383 \\ & \text{CME} + 0.0177 \text{ CMA} - 0.0358 \text{ SAA} + 0.117 \text{ IM} + 0.00684 \text{ EAG} + 0.0448 \text{ CT} \\ & - 0.0444 \text{ SND} + 0.0668 \text{ SE} + 0.0102 \text{ SS.} \end{aligned}$$

El modelo presentado para la clasificación identifica las variables explicativas que ayudan a crear uno en que se seleccionen las variables que puedan influir en la respuesta, descartando aquellas que no aporten información; por ello, en la ecuación anterior se observa que algunos β_i tienen un valor muy bajo con respecto a los demás y que multiplican a las X (tipo de actividad económica); esto sugiere que algunas variables X no aportan significativamente para explicar el valor de Y , por lo que se procede al análisis de resultados.

Tabla 2. *Análisis de variables significativas de la primera regresión lineal múltiple de PO vs UE*

<i>Predictor</i>	<i>Coef.</i>	<i>Coef. de EE</i>	<i>T</i>	<i>P</i>
Constante	0.35133	0.02687	13.08	0.000
SCD	0.00421	0.02587	0.16	0.871
TCA	-0.01931	0.01171	-1.65	0.102
SIB	0.00699	0.02206	0.32	0.752
IMM	-0.03528	0.01569	-2.25	0.026
CME	0.38302	0.04394	8.72	0.000
CMA	0.01770	0.01956	0.90	0.367
SAA	-0.03579	0.04244	-0.84	0.401
IM	0.11651	0.03224	3.61	0.000
EAG	.006843	0.005877	1.16	0.247
CT	0.04479	0.01588	2.82	0.006
SND	-0.04439	0.02381	-1.86	0.065
SE	0.06676	0.03211	2.08	0.040
SS	0.01015	0.01761	0.58	0.565

Fuente: elaboración propia.

De la tabla anterior se concluye que, con un nivel de confianza del 95 % ($1-\alpha$) y un α de .05, los valores de las X de comercio al por menor (CME), industria manufacturera (IM), construcción (CT), industria de medios masivos de comunicación (IMM), y sector educativo (SE) son estadísticamente significativos para la variable regresora, con base en la evidencia del *p-value*; al evaluar el modelo, se observan los siguientes resultados:

Tabla 3. *Resultados del modelo de regresión lineal múltiple completo*

$S = 0.178348$	$R^2 = 78.6 \%$	$R^2 \text{ (ajustado)} = 76.4 \%$
----------------	-----------------	------------------------------------

Fuente: elaboración propia.

La tabla 3 muestra el coeficiente de determinación R^2 , que es una cantidad que puede interpretarse como un factor (porcentaje) de reducción de la variabilidad (incertidumbre) cuando son conocidas las variables independientes, es decir, la proporción de variabilidad total de la variable dependiente respecto a su media que es explicada por el modelo de regresión; en otras palabras, cuanto más se acerque a uno, más poder explicativo tendrá el modelo.

El valor obtenido de 78.6 % infiere un modelo bueno dado que, si el valor de $|R^2|$ está entre 0.5 y 0.85, se dice que es un modelo representativo. Finalmente, en el análisis de varianza del modelo se observan los siguientes resultados:

Tabla 4. Cuadro anova; análisis de varianza del primer modelo de regresión lineal múltiple

Análisis de varianza					
Fuente	GL	SC	MC	F	P
Regresión	13	14.2896	1.0992	34.56	0.000
Error residual	122	3.8806	0.0318		
Total	135	18.1702			

Fuente: elaboración propia.

Se evidencia que el valor P^1 , al ser cercano a 0.000, determina que el modelo es significativo y existe evidencia estadística para rechazar la H_0 y aceptar con la H_a para afirmar que $\beta_j \neq 0$ para al menos una j y por lo tanto X_j si influye en Y .

El modelo presentado explica las variables que aportan al comportamiento de las personas ocupadas o empleadas y, para ajustar aún más el modelo predictivo de regresión, se vuelve a correr el modelo sólo con las 5 variables X (CME, IM, CT, IMM, y SE) que resultaron más significativas y se observa el comportamiento en el nivel del coeficiente de determinación; por lo tanto, la ecuación de la regresión queda de la siguiente manera:

Ecuación 3. Modelo de regresión lineal múltiple ajustado con las UE significativas

$$PO = 0.352 + 0.0317 CT + 0.117 IM + 0.0550 SE + 0.370 CME - 0.0512 IMM$$

Tabla 5. Análisis de variables significativas de la segunda regresión lineal múltiple de PO vs UE

Predictor	Coef.	Coef. de EE	T	P
Constante	0.35160	0.02565	13.71	0.000
IMM	-0.05118	0.01019	-5.02	0.000
CME	0.36964	0.04002	9.24	0.000
IM	0.11659	0.02882	4.05	0.000
CT	0.03174	0.01310	2.42	0.017
SE	0.05504	0.02256	2.44	0.016

Fuente: elaboración propia.

¹ También conocido como *P-value* o valor de probabilidad, se refiere a la probabilidad del estadístico de ser posible bajo la hipótesis nula, siempre y cuando se cumpla la condición de ser menor al nivel de significancia impuesto para rechazar la hipótesis nula.

De la tabla anterior se concluye, con un nivel de confianza del 95 % $(1-\alpha)$ % y un α de .05, que todos los valores de las X son estadísticamente significativos para la variable regresora, con base en la evidencia del *p-value*.

Tabla 6. Resultados del segundo modelo de regresión lineal múltiple

$S = 0.1772$	$R^2 = 77.5 \%$	$R^2 \text{ (ajustado)} = 76.6 \%$
--------------	-----------------	------------------------------------

Fuente: elaboración propia.

En cuanto a los resultados del modelo, se observa que la varianza disminuyó en una cantidad pequeña al reducir las variables independientes y, aunque el valor de ajuste del modelo es similar al anterior, se evidencia que las cinco variables, siguen siendo significativas, lo que hace que el modelo se exprese con menor cantidad de variables determinando la función de la ciudad de Xalapa validado por el ANOVA, que posteriormente se puede ajustar atendiendo al análisis de los residuos presentados, los cuales pueden o no subir este coeficiente para obtener una mejor expresión del modelo.

Tabla 7. Cuadro ANOVA; análisis de varianza del segundo modelo de regresión lineal

Análisis de varianza					
Fuente	GL	SC	MC	F	P
Regresión	5	14.0844	2.8169	89.62	0.000
Error residual	130	4.0859	0.0314		
Total	135	18.1702			

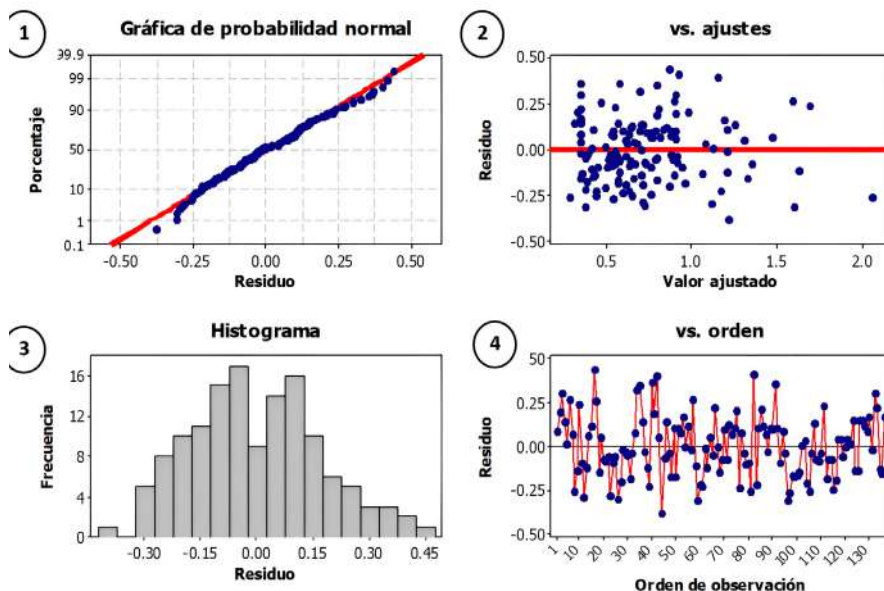
Fuente: elaboración propia.

Análisis de residuos para validar la clasificación

Como lo refiere Cuadras (2007), un estudio más completo del MRLM debe incluir un análisis gráfico de los residuos, efectos de la colinealidad, mínimos cuadrados ponderados, errores correlacionados, selección de las variables, etc., y para ello deben analizarse los residuos de manera gráfica como

en la figura 4. Este es un procedimiento que apoya la identificación de estabilidad del análisis del MRLM y sugiere elementos para lograr un mejor ajuste del modelo.

Figura 4. *Análisis de residuos del modelo de regresión lineal múltiple*



Fuente: elaboración propia.

En la figura 4 se observa, de manera esquemática y descriptiva, el análisis de los residuos que son la diferencia entre los valores observados y los valores estimados, y estos se miden a partir de la recta. La notación matemática de los residuos es la siguiente:

Ecuación 4. *Notación de residuos*

$$\hat{e}_i = Y_i - \hat{Y}$$

Este análisis de residuos aporta elementos significativos para identificar patrones de comportamiento que refuerzan los resultados de las suposiciones teóricas del análisis de varianza (ANOVA).

El gráfico 1 (distribución de probabilidad de la normal) representa el supuesto de normalidad, es decir que los residuos se acercan a la distribu-

ción normal $N(0, \sigma^2)$; sin embargo, se observa también la presencia de algunos valores atípicos que pueden excluirse para determinar si su ausencia permite elevar el coeficiente de determinación.

El gráfico 2 (ajustes) de la figura 4 representa el supuesto de homocedasticidad, es decir que la varianza de los errores de la regresión es la misma para cada una de las observaciones; este comportamiento va ligado a la normalidad, ya que esta presenta media 0 y varianza σ^2 constante en todas las variables independientes. Esto lleva a que en el modelo los coeficientes estimados sean los mejores o más eficientes y, de manera gráfica, no se le atribuye ningún patrón de comportamiento que asemeje heterocedasticidad.

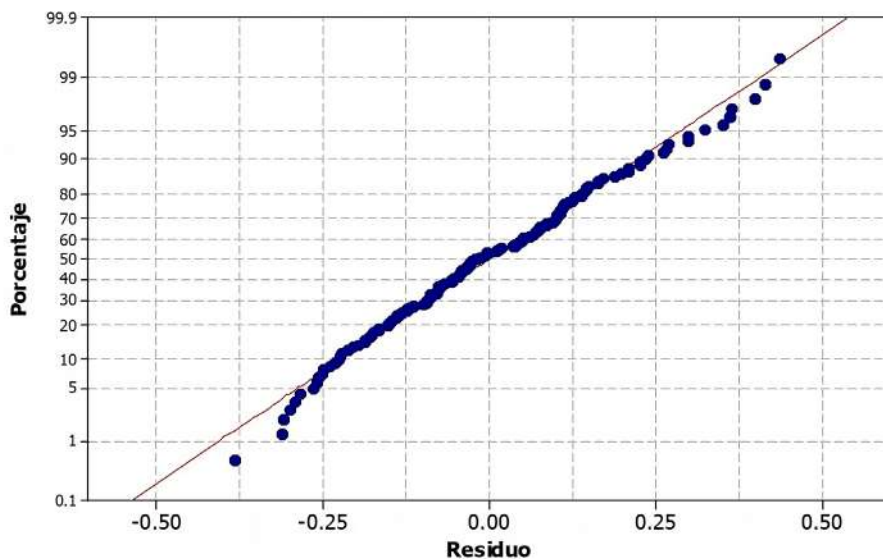
En el gráfico 3 (histograma) de la figura 4 también se presenta, al igual que el número 1, normalidad en los residuos con el primer valor catalogado como dato atípico.

Finalmente, en el gráfico 4 de la figura 4 se observa el supuesto de independencia en el cual, con base en la hipótesis de observaciones imparciales, se entiende que los errores son variables aleatorias independientes y se demuestra por qué no existe un patrón en el gráfico que represente dependencia. Dicho lo anterior, se concluye que, al analizar los residuos del modelo, se presentan los supuestos teóricos de normalidad en residuos, independencia de los errores y homocedasticidad (homogeneidad de varianza).

Retomando el diagrama de normalidad, se había mencionado que se encuentran valores atípicos en la gráfica cuyo análisis debe evidenciar ciertas anomalías o distinciones necesarias en el modelo.

Para detectar el dato u observación que más influye en el coeficiente de determinación se presenta la siguiente tabla, que muestra la salida de Residuos pocos comunes que arroja el programa MINITAB. Además, se utiliza la generalización de la estadística de Cook, la cual permite identificar las observaciones más influyentes en la estimación vía mínimos cuadrados de los parámetros del modelo de regresión lineal múltiple. En el resultado estadístico propuesto por Cook, la influencia de una observación es medida por el cambio en el centro de la región elipsoidal dada cuando la i -ésima observación es eliminada. Para ello, Cook define una medida de distancia D_i que se compara con el percentil de una F con r y $n - r$ grados de libertad; en otras palabras, grandes valores de D_i indican que la observación es influyente (Jiménez, 2001).

Figura 5. Probabilidad normal del segundo MRLM



Fuente: elaboración propia.

Tabla 8. Análisis de observaciones poco comunes o datos atípicos

Observaciones poco comunes			
Observación	Residuo	Residuo estándar	Di Cook
5	0.0119	0.07 X	0.000190
8	0.0646	0.42 X	0.202125
9	-0.2594	-1.73 X	0.168085
10	-0.1410	-1.08 X	0.060860
11	0.2369	1.45 X	0.012266
12	-0.2926	-1.88 X	0.177880
16	0.4354	2.86 RX	0.485846
40	0.3635	2.07 R	0.015307
42	0.3969	2.28 R	0.028195
44	0.3814	-2.29 R	0.118749
56	-0.0229	-0.15 X	0.001137
64	0.0475	0.29 X	0.002755
82	0.4117	2.40 R	0.068982
91	0.3590	2.04 R	0.008355

Fuente: elaboración propia.

En la columna de residuo estándar R denota una observación con un residuo estandarizado grande, mientras que X denota una observación cuyo valor X le concede gran influencia; por ello, con base en los parámetros establecidos en la tabla, la observación 16 tiene la mayor distancia con base en la generalización de Cook, cuyo residuo estandarizado es grande y su valor X concede mucha influencia. Este se representa en el gráfico 4 como el punto localizado en la parte superior; sin embargo, visualmente el dato 44 es el mejor candidato para omitirse a fin de mejorar la modelación de la expresión matemática.

Para verificar si con la exclusión de alguno de estos datos o con ambos es posible aumentar el valor del coeficiente de determinación, se corre el MRLM para los tres casos, con lo que se obtienen los resultados siguientes:

Tabla 9. *Análisis del MRLM omitiendo la observación 16*

$S = 0.172281$		$R^2 = 78.5 \%$		$R^2 \text{ (ajustado)} = 77.7 \%$	
<i>Análisis de varianza</i>					
Fuente	GL	SC	MC	F	P
Regresión	5	14.0159	2.8032	94.44	0.000
Error Residual	129	3.8288	0.0297		
Total	134	17.8447			

Fuente: elaboración propia.

Tabla 10. *Análisis del MRLM omitiendo la observación 44*

$S = 0.174336$	$R2 = 78.4 \%$		$R2 \text{ (ajustado)} = 77.6 \%$		
$Análisis \ de \ varianza$					
Fuente	GL	SC	MC	F	P
Regresión	5	14.2391	2.8478	93.70	0.000
Error Residual	129	3.9207	0.0304		
Total	134	18.1598			

Fuente: elaboración propia.

Tabla 11. *Análisis del MRLM omitiendo las observaciones 16 y 44*

$S = 0.170122$		$R^2 = 79.2 \%$		$R^2 \text{ (ajustado)} = 78.4 \%$	
<i>Análisis de varianza</i>					
Fuente	GL	SC	MC	F	P
Regresión	5	14.1288	2.8258	97.64	0.000
Error Residual	128	3.7045	0.0289		
Total	133	17.8333			

Fuente: elaboración propia.

Según los resultados obtenidos, al omitir las observaciones 16 y 44 es posible mejorar y ajustar un mejor MRLM aumentando el coeficiente de determinación a un 79.2 %; sin embargo, no se van a inhibir estos valores para los siguientes procedimientos bajo el argumento de que el aumento es mínimo y parte de la naturaleza de la estructura urbana de Xalapa es la heterogeneidad de zonas geográficas cuyos valores atípicos representan puntos de inflexión en áreas de influencia o de aglomeración. También se enfatiza en que no se pueden omitir más observaciones porque el MRLM puede perder validez y, en este caso, no se requiere homogeneidad de datos, sino que la heterogeneidad de observaciones permite encontrar sectores, patrones o segregaciones en el comportamiento de la ciudad.

Conclusiones

El estudio de fenómenos urbanos representa una manera de ver lo que acontece en las ciudades, cuyas realidades deben ser explicadas por diferentes metodologías que abonen a la certidumbre de los investigadores. En este trabajo se presentó una manera para determinar la clasificación económica de la ciudad de Xalapa basada en la cantidad de concentración de actividades económicas de diferentes sectores donde la población ocupada se encuentra contratada.

La clasificación económica se determinó con la ayuda de la ciencia de datos, que contribuye a dar certeza al análisis numérico basado en el modelo de regresión lineal múltiple, que permitió generar una ecuación que determina la dependencia de la población económica en 5 sectores importantes: comercio al por menor, construcción, industria manufacturera, información de medios masivos y de información, y sector educativo.

Los sectores económicos pertenecen a la actividad terciaria o de servicios que refleja la historia de las transformaciones de la ciudad de Xalapa, enlazando su proceso histórico como una ciudad de paso entre el puerto de Veracruz y la Ciudad de México. El contraste entre lo que arroja el MRLM con respecto a la historia fundamenta la relación de metodologías que construyen y dialogan para interpretar y explicar los fenómenos urbanos.

La clasificación económica presentada se resume en la ecuación $PO = 0.352 + 0.0317 CT + 0.117 IM + 0.0550 SE + 0.370 CME - 0.0512 IMM$, que no solo reduce la cantidad de sectores con los cuales se explica la población ocupada, sino que representa un 78 % aproximadamente de la explicación total de la ciudad, lo que hace que el coeficiente muestre una alta confiabilidad.

No obstante, la utilización de la ciencia de datos en los fenómenos urbanos también ejemplifica que toda expresión numérica debe evaluarse bajo los parámetros de la estadística para, en primera instancia, asegurar la validez correspondiente y, en segunda, mantener las opciones del investigador para detectar emergencias, anomalías o datos atípicos que pongan en crisis algunos parámetros y lo ayuden a tomar decisiones para mantener los procedimientos y las inferencias con la mejor certeza posible.

Para los usuarios de la estadística, la evaluación de ANOVAS y residuos y la detección de datos atípicos son procedimientos que ayudan a ver con otros ojos las respuestas que necesitan las investigaciones urbanas, no solo para validar los modelos, sino para visualizar las posibles causas y perspectivas que el mismo fenómeno puede ofrecer; en otras palabras, la multiplicidad de caminos para mantener la certeza de explicar el fenómeno con la menor cantidad de elementos estadísticos.

Finalmente, determinar la clasificación de la ciudad de Xalapa permitió encontrar procesos que dialogan interdisciplinariamente con el urbanismo, lo que lleva a mantener ahora la disciplina de las ciencias de la ciudad como una aproximación de la cantidad de aportaciones que otras disciplinas abonan directamente al estudio de fenómenos urbanos. Un diálogo sustancial puede atribuirse al uso de la metodología estadística en la ciencia de datos para dar certeza, confiabilidad y procedimientos robustos a fin de explicar lo que acontece en las ciudades contemporáneas.

Referencias

- Aguilar, A. G., Gratzbord, B., y Sánchez Crispín, Á. (1996). *Las ciudades intermedias y el desarrollo regional en México*. Universidad Nacional Autónoma de México.
- Beaujeu Garnier, J., y Chabot, G. (1972). *Geografía Urbana*. Vences-Vives Editorial.
- Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International Statistical Review / Revue Internationale de Statistique*, 21-26.
- Cuadras, C. M. (2 de febrero de 2007). *Nuevos Métodos de Análisis Multivariante*. CMC Editions.
- Hayashi, Ch. (1998). What is data science? Fundamental concepts and a heuristic example. En Ch. Hayashi, K. Yajima, H.-H. Bock, N. Ohsumi, Y. Tanaka y Y. Baba (Eds.), *Data science, classification, and related methods* (pp. 40-51). Springer. https://doi.org/10.1007/978-4-431-65950-1_3
- INEGI. (2020). *Clasificación para Actividades Económicas*.
- Jiménez M., J. A. (2001). Una Generalización de la Estadística de Cook. *Revista Colombiana de Estadística*, 24 (2), 111-120.
- Johnson, J. (1974). *Geografía Urbana*. Oikos_tau.
- Naur, P. (1974). *Concise survey of computer methods*. Petrocelli Books.
- Naciones Unidas. (2024). <https://www.un.org/sustainabledevelopment/es/cities/>
- Organización Internacional del Trabajo. (2024). <https://www.ilo.org/es/departamento-de-politicas-sectoriales-sector/actividades/temas/economia-urbana>
- Ojeda Ramírez, M. M. (2000). *Modelación de Regresión*. Universidad Veracruzana.
- Ojeda Ramírez, M. M. (2000). *Una Introducción a los Métodos Multivariados*. Universidad Veracruzana.
- Pownall, L. (1956). The Origins of Towns in New Zealand. *New Zealand Geographer*, 115-112.
- Sánchez García, J. A. (Octubre de 2021). Universidad Veracruzana. <https://cdigital.uv.mx/bitstream/handle/1944/51970/SanchezGarciaJuan.pdf?sequence=1&isAllowed=y>
- Tukey, J. (1962). The Future of Data Analysis. *Annals of mathematical statistics*, 1-67.
- Tukey, J. (1977). *Exploratory Data Analysis*. Pearson.
- Zárate Martín, M. A. (2012). *Geografía Urbana; Dinámicas locales, procesos globales*. Centro de Estudios Ramón Areces.