



ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

**INGRESO, POBREZA Y DESIGUALDAD. UN ANÁLISIS
ESPACIO TEMPORAL USANDO APRENDIZAJE
PROFUNDO (DEEP LEARNING).**

**PREDICCIÓN DEL INGRESO A NIVEL CANTONAL EN EL
ECUADOR**

**TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO
REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO
MATEMÁTICO**

JORGE RICARDO SOSA DONOSO

jorge.sosa01@epn.edu.ec

DIRECTOR: ADRIANA UQUILLAS ANDRADE

adriana.uquillas@epn.edu.ec

DMQ, MARZO 2023

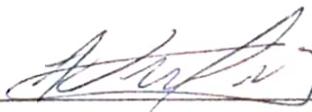
CERTIFICACIONES

Yo, JORGE RICARDO SOSA DONOSO, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.



JORGE RICARDO SOSA DONOSO

Certifico que el presente trabajo de integración curricular fue desarrollado por JORGE RICARDO SOSA DONOSO, bajo mi supervisión.



ADRIANA UQUILLAS ANDRADE
DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el(los) producto(s) resultante(s) del mismo, es(son) público(s) y estará(n) a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

JORGE RICARDO SOSA DONOSO

ADRIANA UQUILLAS ANDRADE

RESUMEN

El ingreso es, sino la más importante, una de las variables más importantes para realizar diversos estudios sobre pobreza y desigualdad. En Ecuador, el medio por el cual se puede acceder a información sobre ingresos de la población es mediante la Encuesta Nacional de Empleo Desempleo y Subempleo (ENEMDU). Sin embargo, esta encuesta no posee representatividad para el nivel desagregado cantonal a pesar de que este sea el segundo nivel más importante en la división administrativa del país. En el presente trabajo se realiza el cálculo del ingreso promedio per cápita para los cantones de Ecuador en los años 2019, 2020 y 2021, en donde se explora el uso de un método no paramétrico llamado Regresión Adaptativa Multivariante por Splines (MARS por sus siglas en inglés) pues da la facilidad de trabajar con interacción entre las variables además de generalizar el modelo de regresión lineal. Además, dependiendo del comportamiento de los datos, se obtiene un modelo con más o menos interacciones; por ejemplo, para el año 2019 se encontró un modelo con interacción de grado 3, así como para el año 2020 bastó un modelo aditivo.

La estimación de los ingresos ayudó para realizar un análisis de autocorrelación espacial evidenciando patrones de asociación de cantones. En la Amazonia se destaca el clúster Bajo-Bajo formado en el año 2020 indicando que si un cantón tiene ingresos bajos sus vecinos también lo tendrán. Este patrón persiste en su mayoría para el año 2021, siendo posiblemente un efecto de la pandemia COVID-19.

Palabras clave: Ingreso, MARS, ENEMDU, Autocorrelación Espacial.

ABSTRACT

Income is, if not the most important, one of the most important variables for various studies on poverty and inequality. In Ecuador, information on the population's income can be accessed through the National Survey of Employment, Unemployment and Underemployment (ENEMDU). However, this survey is not representative at the disaggregated cantonal level, even though this is the second most important level in the administrative division of the country. In this paper we calculate the average per capita income for the cantons of Ecuador for the years 2019, 2020 and 2021, where we explore the use of a non-parametric method called Multivariate Adaptive Regression by Splines (MARS) because it provides the facility to work with interaction between variables in addition to generalizing the linear regression model. Furthermore, depending on the behavior of the data, a model with more or less interactions is obtained; for example, for the year 2019 a model with degree 3 interaction was found, as well as for the year 2020 an additive model was sufficient.

The estimation of income helped to perform a spatial autocorrelation analysis, showing patterns of association of cantons. In the Amazon, the Low-Low cluster formed in the year 2020 stands out, indicating that if a canton has low income, its neighbors will also have low income. This pattern persists for the most part for the year 2021, possibly being an effect of the COVID-19 pandemic.

Keywords: Income, MARS, ENEMDU, Spatial Autocorrelation.

Índice general

1. Descripción del componente desarrollado	1
1.1. Objetivo general	3
1.2. Objetivos específicos	3
1.3. Alcance	4
1.4. Marco teórico	4
1.4.1. Modelo Multivariate Adaptive Regression Spline (MARS)	5
1.4.2. Autocorrelación Espacial	7
2. Metodología	10
3. Resultados, conclusiones y recomendaciones	15
3.1. Resultados	15
3.1.1. Modelo MARS para el año 2019	17
3.1.2. Modelo MARS para el año 2020	20
3.1.3. Análisis de Autocorrelación Espacial	22
3.2. Conclusiones y recomendaciones	26
A. Anexos	29
A.1. Codificación de provincias y cantones de Ecuador	29
A.2. Cantones con muestra representativa para los años 2020 y 2021	35

A.3. Importancia de las variables del modelo MARS y gráficos de dispersión para muestra de entrenamiento y validación para el año 2020.	36
A.4. Informes de autocorrelación espacial y mapas de agrupamiento clúster para los años 2020 y 2021.	38
Bibliografía	40

Índice de figuras

1.1. Representación de la medida de contigüidad de las unidades espaciales para los criterios Torre, Alfíl y Reina.	8
2.1. Valor Agregado Bruto cantonal para el año 2019.	11
2.2. Valor Agregado Bruto no Petrolero cantonal para el año 2019.	12
3.1. Cantones representativos para el año 2019.	16
3.2. Gráfico QQ de los residuos del modelo seleccionado MARS del año 2019.	18
3.3. Importancia de las variables basado en el impacto de los valores de GCV y RSS para el año 2019.	18
3.4. Valores reales y predicho para la muestra de entrenamiento (train) para el año 2019.	19
3.5. Valores reales y predicho para la muestra de validación (test) para el año 2019.	19
3.6. Gráfico QQ de los residuos del modelo seleccionado MARS del año 2020.	20
3.7. Valores reales y predichos para los cantones representativos del año 2021.	21
3.8. Mapa de calor para los cantones de Ecuador para los años 2019 (a), 2020 (b) y 2021 (c).	22
3.9. Informe de Autocorrelación Espacial para el año 2019.	23

3.10 Representación de clúster formados a partir del índice LISA para el año 2019 (a), 2020 (b) y 2021 (c).	24
A.1. Cantones con muestra representativa año 2020.	35
A.2. Cantones con muestra representativa año 2021.	35
A.3. Importancia de las variables basado en el impacto de los valores de GCV y RSS para el año 2020.	36
A.4. Valores reales y predicho para la muestra de entrenamiento (train) para el año 2020.	37
A.5. Valores reales y predicho para la muestra de entrenamiento (test) para el año 2020.	37
A.6. Informe de Autocorrelación Espacial para el año 2020	38
A.7. Informe de Autocorrelación Espacial para el año 2021	39

Capítulo 1

Descripción del componente desarrollado

Un asunto de extrema importancia entre los países de América Latina y en concreto en Ecuador ha sido la lucha contra la pobreza y la desigualdad en sus habitantes. Un indicador que es utilizado comúnmente para la desigualdad es el coeficiente de Gini [28]. Para Ecuador, la brecha de desigualdad desde 2004 hasta 2014 fue en descenso, pero desde entonces no ha podido disminuir su valor manteniéndose alrededor de 0.46 [31][15]. Hecho que debe preocupar al gobierno y que a su vez sirve de incentivo para generar políticas económicas que ayuden en el desarrollo del país, siendo las predicciones macroeconómicas un paso fundamental para lograr este cometido [24]. En este contexto, la principal variable en estudios de pobreza y condiciones de vida es el ingreso ya que es el medio por el cual se puede acceder a los bienes y servicios básicos que poseen individuos, familias, hogares, etc [5].

Contar con información actualizada permite idear planes de desarrollo, planificar y tomar decisiones de manera efectiva; además, focalizar acciones en las necesidades específicas de una zona en particular de la población, puede evitar que afecten a toda la comunidad. Por ello, trabajar con datos a nivel desagregado es de utilidad, pero a su vez, complicado por la cantidad de recursos que conlleva. Una alternativa es trabajar con una encuesta que tenga representatividad a un nivel de desagregación mayor y realizar lo que se conoce como estimaciones en áreas pequeñas

[35], por ejemplo, se ha aplicado este método para determinar la desigualdad por consumo en Ecuador a nivel cantonal y parroquial [29] para el año 2014. Esta metodología depende de la calidad de la información adicional que se pueda complementar que suele venir generalmente de censos al proporcionar información muy vasta y variada; sin embargo el último censo registrado en Ecuador fue en el año 2010.

Actualmente, el medio de recolección de datos de mayor frecuencia temporal en Ecuador es la Encuesta Nacional de Empleo Desempleo y Subempleo (ENEMDU), cuya finalidad es proporcionar información sobre la población económicamente activa, población económicamente inactiva, el empleo, subempleo y desempleo en Ecuador [36]. Su levantamiento de datos empezó trimestralmente a partir de 2008 y en años posteriores se han implementado otras periodicidades (mensual y anual). A pesar de que la frecuencia temporal de ENEMDU es de gran ayuda para la obtención de información actualizada, la representatividad de la encuesta es a nivel nacional y dependiendo de la periodicidad se tiene representatividad a nivel provincial (encuesta anual) o a nivel de zonas urbano, rural y de las 5 ciudades principales del país (Quito, Guayaquil, Cuenca, Ambato y Machala) para la encuesta trimestral, pero no cuenta con representatividad para los cantones.

El presente trabajo intenta mostrar el uso de una herramienta para realizar la estimación del ingreso promedio per cápita por cantón mediante un método no paramétrico que generaliza al modelo de regresión lineal, conocido como Multivariate Adaptive Regression Splines (MARS por sus siglas en inglés) [20]. Este modelo es escogido pues además de brindar la oportunidad de trabajar con un modelo lineal, permite no linealidades e interacciones entre variables [26][43]. La información referente a ingresos será mediante la encuesta ENEMDU; además, se usará las bases de datos del Valor Agregado Bruto (VAB) a nivel cantonal como variables independientes, las cuales, se encuentran en miles de dólares y según el Banco Mundial, es el indicador macroeconómico que mide el valor añadido de cada sector o industria en el proceso productivo de bienes finales [6], siendo también parte del cálculo del Producto Interno Bruto

(PIB). Los datos del VAB se pueden encontrar en las bases de datos del Banco Central del Ecuador [17].

Por otra parte, campos como la criminología [3], medio ambiente [14], epidemiología [41], entre otros, estudian si sus mediciones pueden estar afectadas por el entorno en el que se encuentran, siendo una sospecha que su lugar geográfico incide de una u otra manera. La autocorrelación espacial toma un papel fundamental ya que, mide que tan correlacionada está la característica de interés medida en distintos sitios del espacio, además de ser una causa de que los modelos econométricos no posean, en algunos casos, un buen ajuste; diversos índices y métodos han sido implementados para detectar la autocorrelación espacial con el objetivo de determinar los agrupamientos espaciales de manera general (global) o local [27],[1],[34],[40]. Este análisis se ha implementado para estudiar la relación entre la agrupación espacial de la distribución del ingreso y la desigualdad de las provincias de Argentina [7] o determinar la distribución espacial de desigualdad por ingresos en la unión europea [18]. Por tanto, como un aporte adicional, en este documento se utilizará el índice I de Moran [30] para detectar la existencia o no de autocorrelación espacial en el ingreso promedio per cápita por cantones en Ecuador, este índice ha sido utilizado, por ejemplo, para estudiar la autocorrelación espacial del ausentismo electoral en el país [4].

1.1. Objetivo general

Realizar un modelo de predicción para el ingreso per cápita por cantones en el Ecuador, considerando el modelo MARS.

1.2. Objetivos específicos

1. Entender y homologar las bases de datos de las encuestas ENEMDU.
2. Entender la teoría de los modelos Multivariate Adaptive Regression Splines (MARS).
3. Estructurar la base de datos y construir/transformar variables exóge-

nas.

4. Realizar un análisis de autocorrelación espacial del ingreso per cápita y comparar los resultados obtenidos para distintos periodos.

1.3. Alcance

Dado que no se cuenta con información representativa sobre el ingreso promedio per cápita por cantones en Ecuador, se utilizarán los datos recopilados por la Encuesta Nacional Empleo, Desempleo y Subempleo (ENEMDU) anual, pues su nivel de representatividad es el más cercano a cantón. Con ello, se usarán los cantones que tengan el número de encuestas suficientes para calcular el ingreso promedio per cápita de forma directa y se entrenará un modelo MARS para realizar estimaciones sobre los cantones restantes; se usarán las bases de datos del Valor Agregado Bruto (VAB) cantonal como variables independientes. Se realizará la estimación para los años 2019, 2020 y 2021 con el objetivo de comparar estos periodos y observar si la economía de los cantones en Ecuador se vio afectada por agentes externos como la pandemia COVID-19; resultados que se pueden obtener mediante el análisis clúster espacial mediante el índice de Moran local.

1.4. Marco teórico

A continuación, se expondrán los conceptos teóricos más relevantes de los que se basa este trabajo. La técnica de modelado escogido para este trabajo es mediante un modelo de regresión no paramétrico que puede verse como una generalización del modelo de regresión lineal paso a paso o como una modificación de los árboles de clasificación y regresión (CART por sus siglas en inglés) [26]. Se trata del modelo de regresión adaptativa de spline multivariante (MARS) que se explica a continuación:

1.4.1. Modelo Multivariate Adaptive Regression Spline (MARS)

Supongamos que Y es una variable a explicar, $X = (X_1, \dots, X_p)$ son p variables independientes y además se tiene N valores observados (x_i, y_i) , $i = 1, \dots, N$ con $x_i = (x_{i1}, \dots, x_{ip})$ y y_i la observación i de la variable a explicar; MARS hace uso de funciones lineales por partes de la siguiente forma:

$$(x - t)_+ = \begin{cases} x - t & \text{si } x > t \\ 0 & \text{otro caso} \end{cases}$$

$$(t - x)_+ = \begin{cases} t - x & \text{si } x < t \\ 0 & \text{otro caso} \end{cases}$$

Dichas funciones tienen el signo $+$ como sufijo para indicar que se toma la parte positiva de la resta dentro del paréntesis, adicionalmente a $(t - x)_+$ se suele llamar reflejo de $(x - t)_+$. Como siguiente paso, se crea una colección C de ellas, es decir:

$$C = \{(X_j - t), (t - X_j)\} \quad \begin{array}{l} t \in \{x_1, x_2, \dots, x_N\} \\ j = 1, 2, \dots, p \end{array}$$

A los valores t se los conoce como nodos tomando un valor equivalente a las observaciones de la variable X_j . Las funciones de C tienen como única variable a X_j y t es un valor fijo, haciendo posible adoptar la siguiente notación $h(X_j) = (X_j - t)_+$. En primer lugar, se explicará el modelo más simple (modelo MARS sin interacciones) para luego extenderlo al modelo general:

Se parte de una regresión lineal con pasos hacia adelante (forward), empezando con la función constante $h_0(X) = 1$; en la siguiente iteración ingresan funciones de la forma:

$$\beta_1(X_j - t)_+ + \beta_2(t - X_j)_+ \quad t \in (x_{ij})$$

Si realizamos este procedimiento iterativamente tendremos un modelo equivalente a:

$$f(X) = \beta_0 + \sum_{m=1}^N \beta_m h_m(X)$$

Luego los coeficientes β son estimados con una minimización de la suma de cuadrados del error, es decir, mediante una regresión lineal estándar.

Notemos como en esta primera fase se añaden dos funciones a la vez al modelo, esto con el objetivo de que sobreajuste los datos por lo que una selección hacia atrás es utilizada en una segunda fase, eliminando aquellos que provoquen un incremento menor en el error cuadrático residual en cada etapa y por tanto tenemos un modelo estimado f_λ con λ números de términos. Para más detalles revisar [20]. Para estimar el valor de λ , MARS utiliza la validación cruzada generalizada definida por:

$$GCV(\lambda) = \frac{\sum_{i=1}^N (y_i - \hat{f}_\lambda(x_i))^2}{(1 - M(\lambda)/N)^2}$$

Donde $M(\lambda)$ es el número efectivo de parámetros del modelo, es decir, el número de términos en el modelo y el número de parámetros usados para seleccionar la posición óptima de los nodos.

En el caso general, se añade la posibilidad de tener interacciones entre las variables. Si denotamos a Δ el conjunto de todas las funciones que ingresaron al modelo por el método forward, en cada iteración se considera como una función base a la multiplicación entre una función de Δ y una de C , es decir para la iteración m -ésima añadimos a Δ funciones de la forma:

$$\beta_m h_l(X)(X_j - t)_+ \quad y \quad \beta_{m+1} h_l(X)(t - X_j)_+ \quad t \in (x_{ij}); h_l(X) \in \Delta$$

Se suele pagar un precio de tres parámetros para seleccionar un nodo en una regresión lineal a trozos. Si tenemos r funciones base linealmente independientes y K nodos a ser seleccionados, entonces la fórmula es $M(\lambda) = r + cK$ donde $c = 3$ si el modelo tiene interacciones ($c = 2$ si el modelo es aditivo) [8]

Una restricción que tiene el modelo MARS es no incluir en una misma interacción la misma variable input más de una vez, puesto que se puede modelar de una forma más estable mediante funciones lineales a trozos. Esto se puede evitar fijando un límite en el orden de interacción

1.4.2. Autocorrelación Espacial

La Autocorrelación Espacial (AE), como su nombre lo indica, se encarga en determinar la correlación de una variable consigo misma medida en distintos sitios del espacio geográfico (unidades o zonas espaciales); se puede utilizar para descubrir, comprender o analizar si una unidad espacial en particular se ve afectada por sus vecinos [40], manifestándose en patrones que pueden verse como una de las tres siguientes posibilidades:

1. Autocorrelación espacial positiva (agrupamiento): ocurre cuando valores de la variable en estudio en zonas espaciales vecinas tienden a agruparse mediante valores altos con valores altos o valores bajos con valores bajos.
2. Autocorrelación espacial negativa (dispersión): ocurre cuando los valores de la variable en estudio en zonas espaciales vecinas son distintos y tienden a comportarse de manera dispersa en el espacio.
3. Sin autocorrelación espacial (aleatorio): se da cuando no ocurre ninguno de los dos ítems anteriores.

En este contexto, encontrar una herramienta que permita determinar con qué unidades espaciales tiene relación un lugar en específico se vuelve fundamental. El concepto de matriz de proximidad, contigüidad o de pesos espaciales (W) permite entonces observar la intensidad de la interdependencia entre cada par de zonas espaciales [39]. A pesar de no existir una definición específica para W se puede adoptar como una medida de contigüidad a la representación del movimiento de 3 piezas de ajedrez principalmente: Reina (queen), Torre (rook), Alfil (bishop) [9][2] (Figura 1.1)

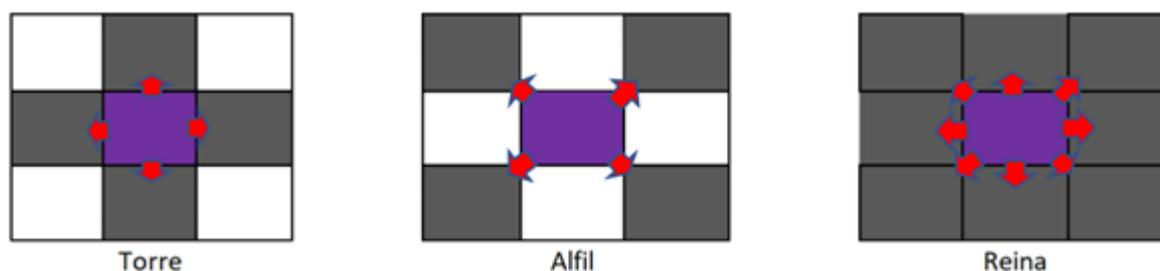


Figura 1.1: Representación de la medida de contigüidad de las unidades espaciales para los criterios Torre, Alfil y Reina.

Para desarrollar el análisis de AE se han propuestos varios índices para contrastar la hipótesis nula de ausencia de autocorrelación espacial, como el índice G de Getis y Ord [23], el índice C de Geary [19][22]; sin embargo, el más utilizado ha sido el índice I de Moran, además de ser uno de los más antiguos, se asemeja al coeficiente de correlación de Pearson [27] al manejar sus valores entre -1 y 1 . El índice está definido de la siguiente manera:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Donde, n es el número de unidades geográficas; w_{ij} es un peso espacial de la matriz W indicando si la zona i y j están próximas; x_i es la característica de interés medida en el sitio i y S_0 es la suma de todos los pesos de la matriz W . Bajo la hipótesis nula de no existencia de autocorrelación espacial, la interpretación del índice es la siguiente: si el valor del índice es cero no existe suficiente evidencia estadística para rechazar la hipótesis nula; caso contrario, si el índice es distinto de cero rechazamos H_0 . Más aún si el índice es mayor a cero (positivo) se tiene correlación espacial positiva; si es menor a cero se tiene correlación espacial negativa. Dado que se realiza una prueba de hipótesis, se puede rechazar H_0 , mediante un valor p o con un valor de la variable estandarizada I (obteniendo una puntuación z) mayor al cuantil de la distribución normal determinado por el error Tipo I (para más detalles observar [1],[30]).

El índice expuesto anteriormente toma en cuenta variables cuantitativas, que es el caso de este trabajo; para determinar asociación espacial para variables cualitativas se puede referir a [37].

El análisis de autocorrelación espacial local permite conocer el comportamiento local de las unidades espaciales que muchas veces pasan desapercibidos al estudiarlos de manera global [40] mediante la formación de clústeres que se forman al tener valores altos/bajos de la variable en estudio. Para el presente trabajo se hará uso del método Local Indicators of Spatial Association (LISA) que descompone el índice I de Moran y verifica el aporte de cada unidad espacial para la formación del índice global permitiendo así que aquellas unidades espaciales que sean significativos pasen a formar parte de los respectivos clústeres [9].

Capítulo 2

Metodología

Ecuador cuenta con 24 provincias que a su vez están divididas en 221 cantones. En este trabajo nos concentramos en Ecuador continental, por tanto, no se consideran los 3 cantones de la provincia de las Galápagos, obteniendo 218 cantones. En el Anexo A.1 se muestra una tabla con el nombre del cantón y su respectivo código el cual será usado para los posteriores análisis. La metodología empleada en el presente trabajo se compone de dos enfoques principales. El primero de ellos es un enfoque cuantitativo analítico, que se detalla a continuación: La variable de estudio es el ingreso promedio per cápita (IPPC) por cantones que se obtendrá mediante la encuesta ENEMDU; sin embargo, esta encuesta no tiene representatividad a este nivel de desagregación, por tanto, se utilizará las proyecciones poblacionales cantonales para los años 2019, 2020 y 2021 [16], así como, a partir de encuestas anteriores se obtendrán las varianzas para el ingreso per cápita de cada cantón para calcular el tamaño de muestra mínimo para tener representatividad a dicho nivel. Una vez realizado este cálculo, se utilizarán aquellos cantones que tengan un número de encuestas igual o mayor al tamaño de muestra mínimo con la finalidad de entrenar un modelo Multivariate Adaptive Regression Splines (MARS) y se estimará la variable dependiente para los cantones que resultaron no tener la suficiente cantidad de encuestas para obtener el valor del ingreso per cápita de manera directa y sobre aquellos cantones que por el diseño muestral de ENEMDU [36] no fue posible la toma de

muestras. Se usarán como variables independientes las industrias que considera el Banco Central en el cálculo del Valor Agregado Bruto. Para Ecuador se miden las siguientes industrias: salud; enseñanza; actividades financieras; transporte, información y comunicaciones; entre otras. En la Figura 2.1 se pueden observar las industrias en las que se mide el VAB destacando la Manufactura y las Actividades profesionales e inmobiliarias como aquellas que tuvieron la mayor aportación al VAB en el año 2019.

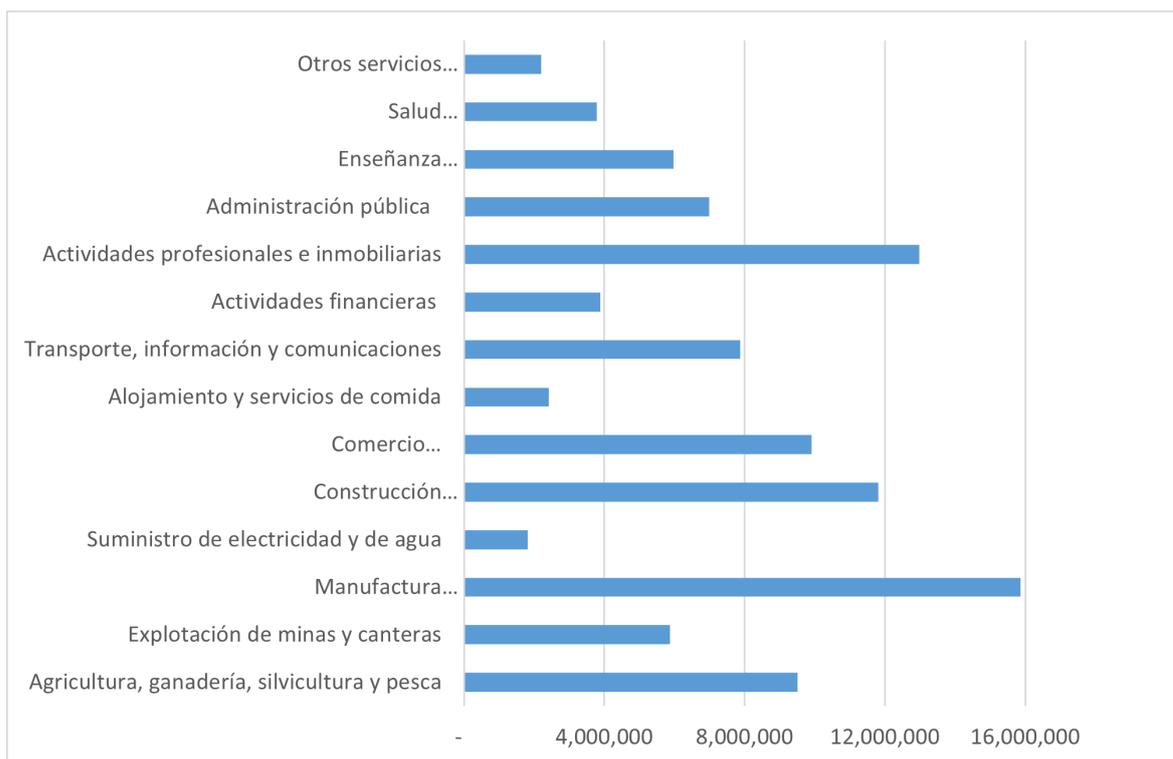


Figura 2.1: Valor Agregado Bruto cantonal para el año 2019.

Así mismo se consideraron VAB no petrolero mediante el consumo intermedio y la producción. (En la Figura 2.2 se observa el VAB no Petrolero para el año 2019)

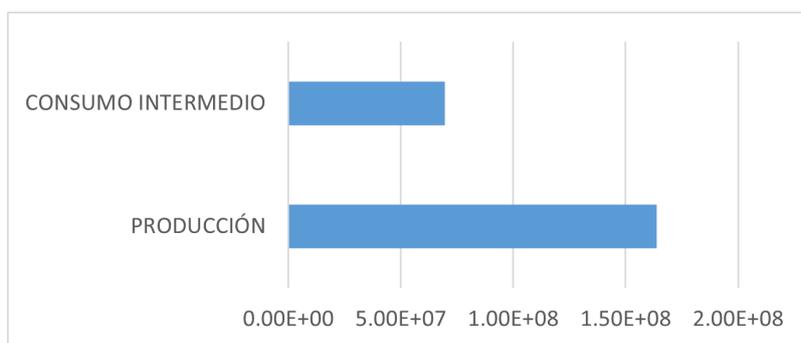


Figura 2.2: Valor Agregado Bruto no Petrolero cantonal para el año 2019.

Para facilitar el tratamiento de las variables independientes se realizó una codificación a cada una (ver cuadro 2.1). Para la fase de modelamiento se considerará el logaritmo tanto de la variable dependiente como las variables independientes:

Codificación	Variable
AASP	Agricultura, ganadería, silvicultura y pesca
EMC	Explotación de minas y canteras
MANU	Manufactura
EA	Suministro de electricidad y de agua
Const	Construcción
ASC	Alojamiento y servicios de comida
TIC	Transporte, información y comunicaciones
AF	Actividades financieras
API	Actividades profesionales e inmobiliarias
ADP	Administración pública
otros	Otros servicios
EcoT	ECONOMÍA TOTAL
Ense	Enseñanza
CI	CONSUMO INTERMEDIO
Comercio	Comercio
PRODUCCION	PRODUCCIÓN
IPPC	Ingreso Promedio Per Cápita

Cuadro 2.1: Codificación de las variables a usar.

Posteriormente se realiza una limpieza de los datos faltantes, como también de datos atípicos mediante la técnica de vecinos más cercanos [25] (KNN por sus siglas en inglés). Para comprobar la eficacia del método se destina el 80% de los datos para una muestra de entrenamiento (train)

y el 20% restante para una muestra de validación (test). Cabe mencionar que se adicionaron dos variables categóricas referente a si el cantón pertenece a una de las tres regiones del país (Costa, Sierra u Oriente), como también, indicando a qué provincia pertenece (en el Anexo A.1 se encuentra la codificación de las provincias de Ecuador).

El procedimiento de entrenar un modelo MARS únicamente será desarrollado para los años 2019 y 2020 ya que al momento de realización de este documento no se cuenta con la información del VAB para el año 2021, por tanto, para dicho año se realizó la estimación en dos fases: En primer lugar, se consideró a los cantones representativos del año 2021 de la encuesta ENEMDU y con la información de 2019 y 2020 se parametriza una media ponderada de la siguiente forma:

$$IPPC_{21,i}^* = \alpha_1 IPPC_{19,i} + \alpha_2 IPPC_{20,i}$$

Con el objetivo de encontrar los mejores α_1 y α_2 que minimicen:

$$error = \sum_{i=1}^n (IPPC_{21,i}^* - IPPC_{21,i})^2$$

Donde:

- $alpha_k$ Parámetros por encontrar con $k = 1, 2$.
- n : Número de cantones representativos de la encuesta ENEMDU para el año 2021.
- $IPPC_{j,i}$: Valores reales del logaritmo del ingreso per cápita para el cantón i del año j con: $j = 19, 20, 21$; $i=1, \dots, n$
- $IPPC_{21,i}^*$: Valor aproximado para el logaritmo del ingreso per cápita para el cantón i del año 2021.

En la fase 2 se realiza una segunda ponderación con la intención de disminuir el error de aproximación de la siguiente manera:

$$IPPC_{21,h,l}^{**} = \alpha_3 IPPC_{21,h,l}^* + \alpha_4 m_{l,20}$$

Donde:

$alpha_k$	Parámetros por encontrar con $k=3,4$.
$IPPC_{21,h,l}^*$	Estimación del logaritmo del ingreso per cápita con los parámetros $alpha_1$ y $alpha_2$ para el cantón h de la provincia l ; con $l = 1, \dots, 23$ y $h = 1, \dots, H_l$
H_l	Cantidad de cantones de la provincia l
$m_{l,20}$	Promedio del logaritmo del IPPC para la provincia l del año 2020
$IPPC_{21,h,l}^{**}$	Valor aproximado para el logaritmo del IPPC del cantón h de la provincia l

Una vez calculado el IPPC para todos los cantones y para los años en estudio, se procede al segundo enfoque que lleva este documento que se trata de un enfoque exploratorio, descriptivo, en donde se realiza un análisis de autocorrelación espacial del ingreso per cápita por cantones. Posteriormente se realizará un análisis de autocorrelación local para determinar los grupos o clústeres donde se indicará como se diferencian o se asemejan cada unidad espacial. Finalmente se realizará la comparación para los distintos periodos en donde se podrá observar si aumentó o disminuyó la agrupación de cantones con ingresos altos o bajos o se agrupan de distinta manera.

Capítulo 3

Resultados, conclusiones y recomendaciones

3.1. Resultados

Para estimar el tamaño de muestra necesaria que haría que la información de ingresos de un cantón sea representativa se consideró muestreo aleatorio simple con un error $e = \$20$ y un nivel de confianza del 95%; sin embargo, el número de cantones con muestra representativa tanto para los años 2020 y 2021 era menor al 15%. Por tanto, se decidió variar el nivel de confianza al 85% obteniendo para los periodos considerados un total de 108, 59 y 45 cantones representativos en 2019, 2020 y 2021, respectivamente. En la Figura 3.1 se puede observar en color rojo los cantones representativos para el año 2019 y en color negro aquellos que no los son (lo propio para los años 2020 y 2021 se visualiza en el Anexo A.2)



Figura 3.1: Cantones representativos para el año 2019.

Como se mencionó anteriormente, al momento de realizar este trabajo la información sobre el VAB únicamente se encontraba disponible para los años 2019 y 2020, por tanto, se van a entrenar dos modelos con la finalidad de particularizar los resultados para cada año, puesto que por la pandemia de COVID-19 existen diferencias en el tamaño de las muestras, además de un impacto económico, lo cual, se describe a continuación.

En primer lugar, debemos realizar una búsqueda de datos faltantes. En el cuadro 3.1 constan las variables que mostraron datos perdidos siendo EMC y Salud las que presentaron un mayor porcentaje de ellos, por tanto, serán eliminadas del estudio. Para los otros casos se utilizará el promedio de su variable como método de imputación.

Variable	% Datos Perdidos
EMC	61.01
Salud	32.11
ASC	2.29
AF	0.46
otros	0.46

Cuadro 3.1: Valores faltantes para las variables independientes para el año 2019.

Como siguiente paso se realiza una búsqueda de datos atípicos mediante la técnica de KNN fijando el valor de $k = 5$, obteniendo un total de

3 datos atípicos correspondientes a la información de los cantones Quito, Guayaquil y Cuyabeno, los cuales fueron separados para continuar con el estudio.

3.1.1. Modelo MARS para el año 2019

Para realizar el modelo MARS se utilizó el software estadístico R mediante la librería earth [21] con la función del mismo nombre. De acuerdo con la revisión teórica, esta función nos permite variar el grado de interacción de las variables, el número máximo de funciones dentro del modelo entre otros. Mediante una búsqueda por validación cruzada con criterio de minimización la raíz del error cuadrático medio (RMSE por sus siglas en inglés), variando los hiperparámetros antes mencionados se determina el mejor modelo con un $R^2 = 0,85$:

$$\begin{aligned}
 IPPC = & 5,508 + 0,192 * h(AF - 8,447) - 0,097 * h(8,447 - AF) \\
 & + 1,351 * prov_{11} * h(7,984 - ADP) - 0,445 * h(Ense - 8,962) \\
 & + 0,343 * h(8,962 - Ense) - 0,24 * prov_{19} * h(8,447 - AF) \\
 & + 0,572 * h(8,447 - AF) * regionSierra + 0,266 * h(PRODUCCION - 12348) \\
 & - 0,135 * h(9,13863 - Comercio) - 1,173 * h(7,24887 - EA) * h(Ense - 8,96243) \\
 & - 0,09 * Const * h(8,44739 - AF) * regionsierra \\
 & + 0,198 * h(8,85288 - API) * h(Ense - 8,96243)
 \end{aligned}$$

La Figura 3.2 muestra que los residuos de este modelo se distribuyen normalmente donde además las muestras (13, 22 y 46) son datos atípicos detectados que fueron excluidos al momento de realizar el método backward.

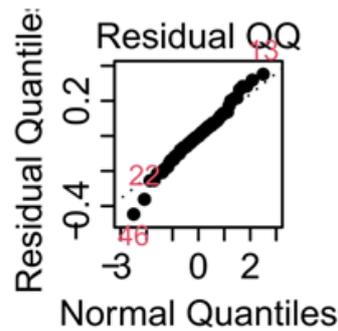


Figura 3.2: Gráfico QQ de los residuos del modelo seleccionado MARS del año 2019.

Debido a que el modelo MARS posee un proceso de poda para determinar las mejores variables que realizan una reducción en la estimación del error de GCV; esto nos indica una medida de importancia también sobre estas variables, siendo aquellas que siempre fueron escogidas con un nivel de importancia del 100% y aquellos predictores que nunca ingresaron al modelo con un nivel de importancia de cero. En la Figura 3.3 se muestra la importancia de las variables para el modelo descrito anteriormente, destacando a Actividades Financieras (AF) como la más influyente y Comercio como la menos influyente (los predictores que no se muestran en la figura poseen un nivel de importancia de cero). Alternativamente, podemos observar también cambios en las sumas residuales de cuadrados (RSS por sus siglas en inglés) obteniendo resultados muy similares.

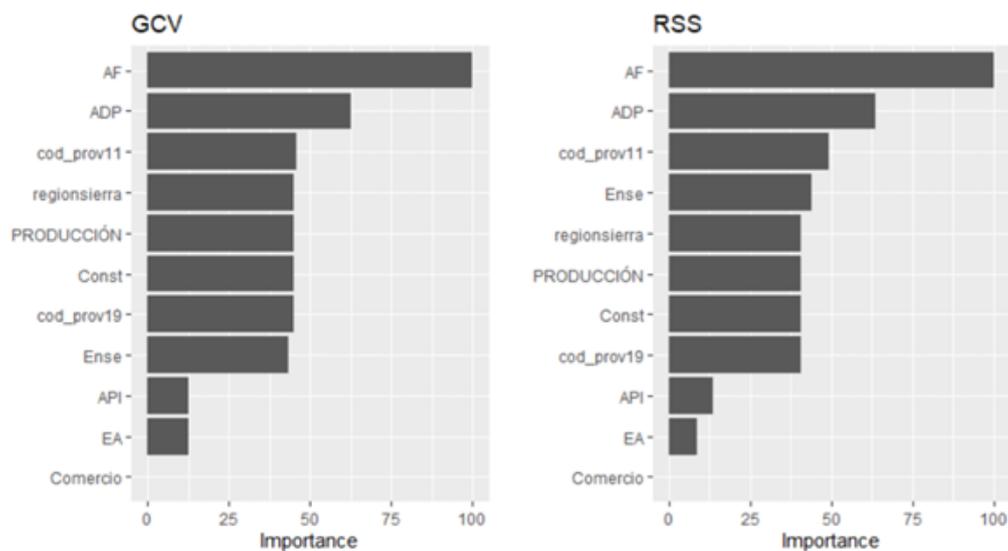


Figura 3.3: Importancia de las variables basado en el impacto de los valores de GCV y RSS para el año 2019.

Los valores reales (eje horizontal) y predichos (eje vertical) se muestran en la Figura 3.4 para la muestra de entrenamiento del año 2019. Notemos que no existen grandes diferencias entre ellos, de igual forma para la muestra de validación (Figura 3.5).

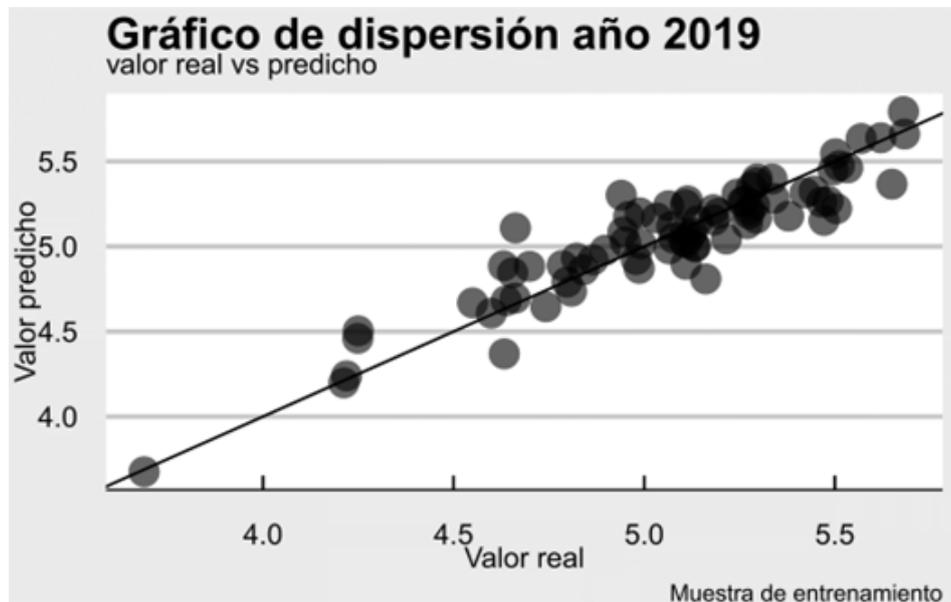


Figura 3.4: Valores reales y predicho para la muestra de entrenamiento (train) para el año 2019.

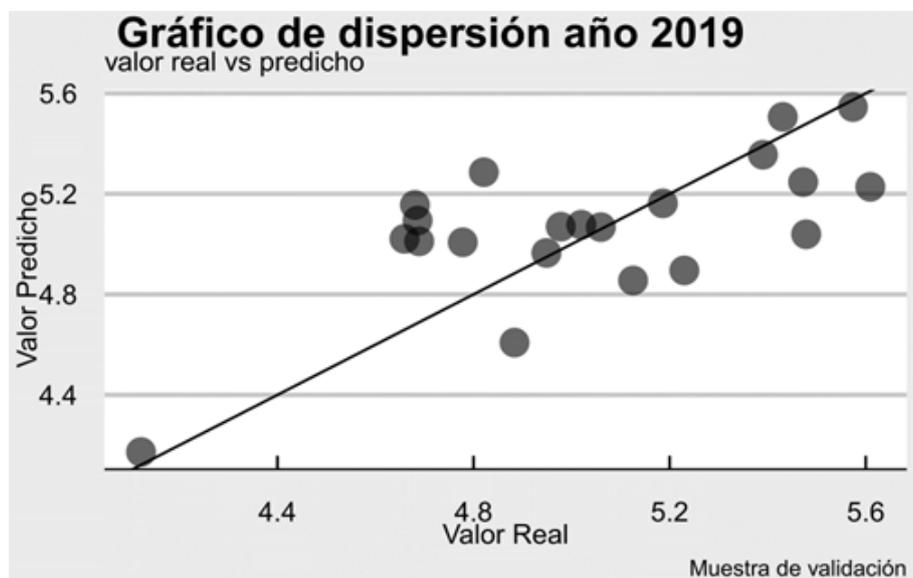


Figura 3.5: Valores reales y predicho para la muestra de validación (test) para el año 2019.

3.1.2. Modelo MARS para el año 2020

Para el año 2020, es válido mencionar que, por la pandemia las encuestas de ENEMDU fueron realizadas de manera telefónica lo que causó que no se tenga confiabilidad en los datos como también una disminución en el tamaño de muestra a comparación con el año anterior. Con un procedimiento similar a lo antes desarrollado, el mejor modelo que se encontró tiene $R^2 = 0,64$, resultando ser un modelo aditivo sin interacciones entre las variables.

$$IPPC = 4,609 - 0,097h(5,655 - AF) - 0,964 * prov_{16} + h(9,238 - TIC) \\ - 0,484 * h(7,67 - API) + 0,373 * h(AASP - 10,441) + 0,158 * h(6,299 - otros)$$

La Figura 3.6 muestra que los residuos de este modelo se pueden asumir que se distribuyen normalmente donde además las muestras (7, 30 y 36) son datos atípicos detectados que fueron excluidos al momento de realizar el método backward.

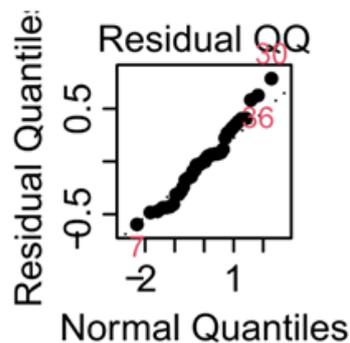


Figura 3.6: Gráfico QQ de los residuos del modelo seleccionado MARS del año 2020.

De nueva cuenta, para la importancia de las variables, Actividad Financiera (AF) es la que sobresale y otros servicios (otros) es la de menor aporte; además, se tiene un comportamiento similar para los dos tipos de estimaciones del GCV y RSS (Anexo A.3) Los valores reales y predichos muestran gran similitud al igual que en el modelo realizado para el año 2019 (Anexo A.3)

Como se comentó en la sección anterior, para el año 2021 se realiza la estimación en dos fases. En la fase uno se pudo determinar que $\alpha_1 = 0,51$

y $\alpha_2 = 0,49$ obteniendo un error=6.69 En la fase dos se intentó disminuir este error encontrando que $\alpha_3 = 0,81$ y $\alpha_4 = 0,19$ con una reducción del error en 0,33. En la Figura 3.7 se muestran los valores reales y predichos utilizando esta metodología para los cantones con muestra representativa del año 2021.

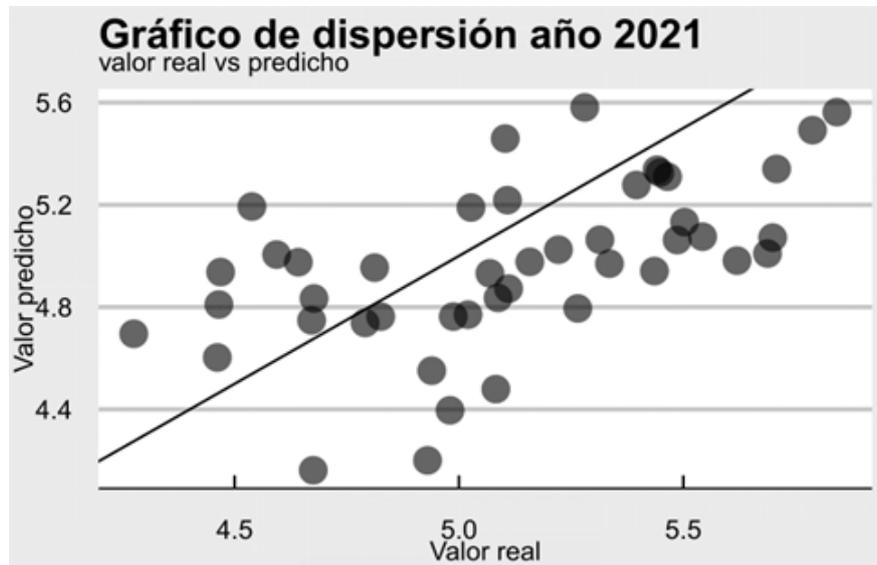


Figura 3.7: Valores reales y predichos para los cantones representativos del año 2021.

3.1.3. Análisis de Autocorrelación Espacial

En la sección anterior, mediante los modelos desarrollados, se obtuvo los valores del logaritmo del ingreso promedio per cápita para todos los cantones de Ecuador haciendo posible comenzar con la segunda fase de este trabajo mediante el análisis de autocorrelación espacial a nivel cantonal. En primer lugar, dado que la relación de orden entre la variable dependiente y su logaritmo se mantiene, se decidió seguir trabajando con el logaritmo por facilidad; por tanto, la Figura 3.8 nos muestra el mapa de calor del logaritmo del ingreso per cápita para el año 2019 (a), 2020 (b) y 2021 (c) donde podemos apreciar que valores altos muestran valores altos a su alrededor y de igual manera para los valores bajos. Estos comportamientos dan un primer vistazo a la existencia de autocorrelación espacial.

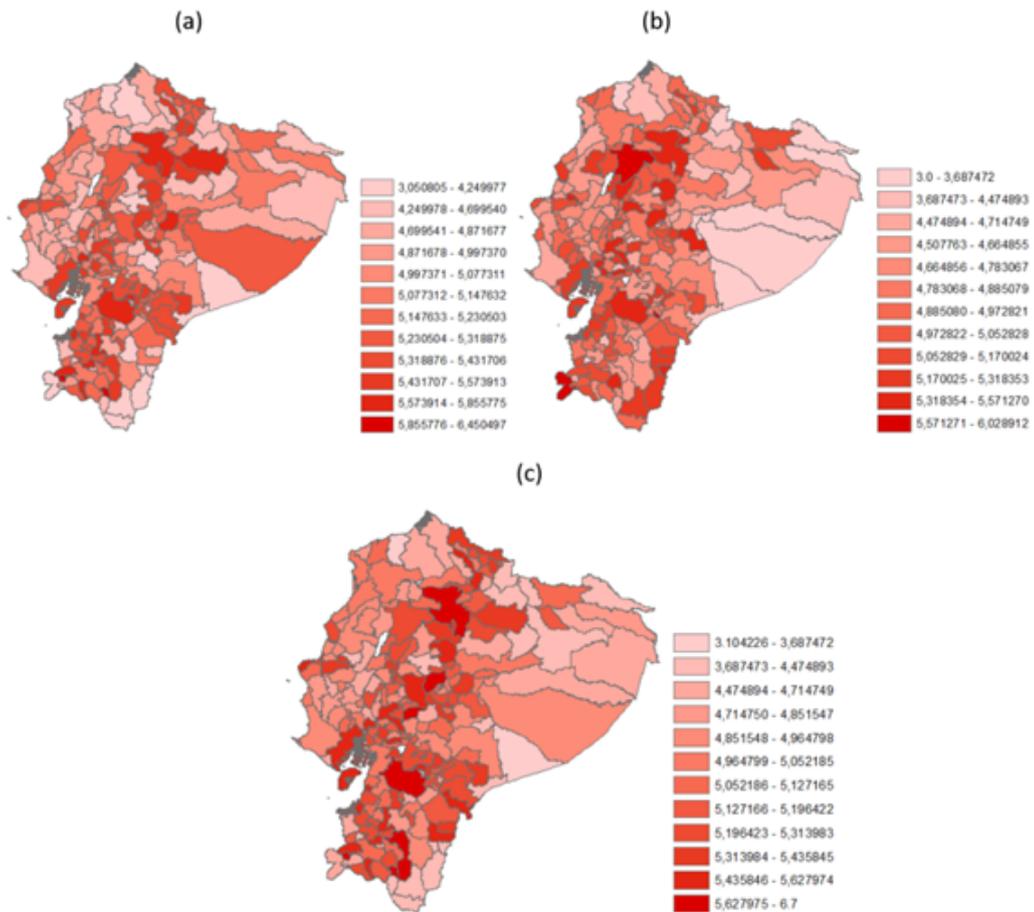


Figura 3.8: Mapa de calor para los cantones de Ecuador para los años 2019 (a), 2020 (b) y 2021 (c).

Como se mencionó en la parte teórica, el índice I de Morán permite determinar si un conjunto de datos está auto correlacionado espacialmente o no, midiendo su significancia mediante una puntuación z y un valor p . Para desarrollar el índice de Moran se utilizó el software ArcMap. Para el año 2019, se indica un valor $z > 2,5$ y un valor $p < 0,01$, es decir, es poco probable que el patrón de distribución espacial sea aleatorio; sin embargo, el valor del índice refleja una correlación positiva débil. De igual manera para los años 2020 y 2021 la probabilidad de que la distribución espacial sea aleatoria es menor al 1%. El reporte de los resultados nos muestra también cual es el patrón en el que se manifiesta la autocorrelación espacial, siendo el agrupamiento el que se manifiesta en todos los casos. A continuación, se evidencia el informe de autocorrelación espacial para el año 2019 (Índice de Moran global, Figura 3.9), lo propio se muestra en el Anexo A.4 para los años 2020 y 2021

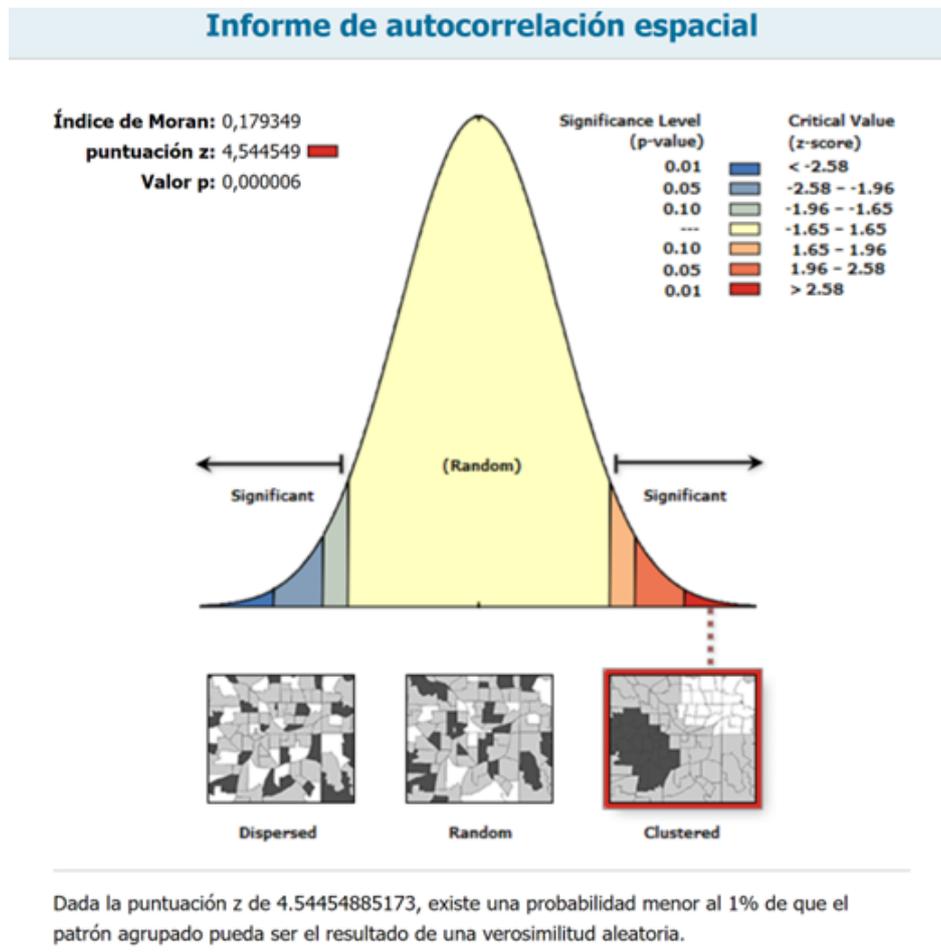


Figura 3.9: Informe de Autocorrelación Espacial para el año 2019.

Al aplicar el índice de Moran de manera local para los años en estudio, se evidencia la existencia de grupos de cantones con distintas características:

1. Bajo-Bajo: si un cantón tiene ingresos bajos, se espera que sus vecinos también tengan ingresos bajos.
2. Bajo-Alto: si un cantón tiene un ingreso bajo se espera que sus vecinos tengan ingresos altos.
3. Alto-Bajo: si un cantón tiene un ingreso alto, se espera que sus vecinos tengan ingresos bajos.
4. Alto-Alto: si un cantón tiene ingresos altos, sus vecinos también tienen ingresos altos.

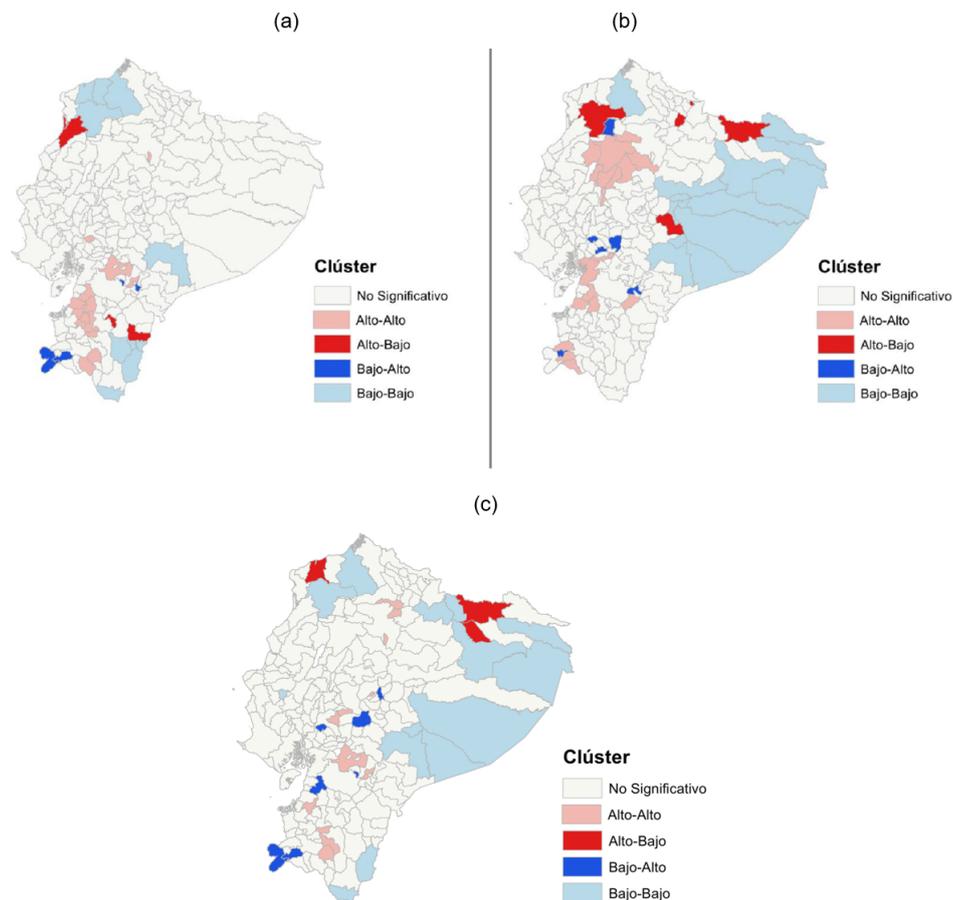


Figura 3.10: Representación de clúster formados a partir del índice LISA para el año 2019 (a), 2020 (b) y 2021 (c).

De la Figura 3.10 podemos destacar los clústeres Alto-Alto de los años 2019 y 2020; el primero formado, en su mayoría, por cantones de la provincia del Oro, hecho que no debe sorprender pues cuenta con una de las cinco ciudades más importantes del país como es Machala, y como lo dice en [32] es una de las provincias que cuenta con grandes posibilidades de crecer económicamente brindando mejor calidad de vida a sus ciudadanos. De igual forma, en el año 2020 se presenta el clúster formado por cantones alrededor de la ciudad Santo Domingo. A pesar de que aproximadamente en marzo de dicho año se presentan los primeros casos relacionados con la pandemia COVID-19 en Ecuador; el comercio, en la cual se fundamenta gran parte de la economía local, repuntó y en el mes de junio tuvo niveles de ventas similares con respecto a dicho mes del año anterior, hecho que entre otros factores se debe a la habilitación del transporte terrestre inter cantonal que ayudó a conectar a este cantón con sus vecinos [13].

Sin embargo, a partir del año 2020 el clúster Bajo-Bajo se presenta en su mayoría por cantones de la Amazonia; esta realidad se mantuvo para el año 2021. A pesar de que los pueblos amazónicos han mejorado su calidad socioeconómica con políticas que vienen del gobierno, mejorando el acceso a servicios básicos, donde un punto clave fue la inclusión de empresas petroleras o mineras [38] al dar trabajo a la población local; la llegada del covid-19 deterioró la situación económica de este sector puesto que muchas empresas al no cumplir los protocolos de protección contra el virus, realizaron rotaciones del personal con gente foránea, lo que ocasionó un incremento en el número de casos registrados en la población local e incluso se llegó a tener fallecidos [42]; con ello, varios pueblos indígenas exigieron la salida de estas empresas de su territorio [10]. Por otra parte, las condiciones geográficas del lugar suponen altos costos de movilización para acceder a centros de salud o así también la población joven mediante su derecho a la educación con la nueva modalidad virtual donde tanto profesores como alumnos deben trasladarse para contar con los servicios de internet lo que supone también, un alto riesgo de contagio [42]. Cosa similar ocurre para los cantones de la provincia de Esmeraldas donde, tanto para el año 2019 y 2021 aparecen bajo la etiqueta Bajo-Bajo sumando además que en esta zona existe alta tasa de criminalidad y que además es la provincia más pobre de la costa

ecuatoriana [33].

Además, se presentan cantones bajo la etiqueta Bajo-Alto que se encuentran en la zona de la sierra sur en su mayoría; así como, cantones identificados bajo el patrón Alto-Bajo. Finalmente, se observa que se tienen algunos patrones aislados que son estadísticamente significativos, es decir, el índice de Moran Local es diferente de cero, sin embargo, para algunos de sus vecinos es estadísticamente cero.

3.2. Conclusiones y recomendaciones

En el presente trabajo se presentó una alternativa para poder calcular el ingreso promedio per cápita sobre los cantones de Ecuador mediante un modelo no paramétrico basado en splines multivariante de regresión adaptativa (MARS). Se pueden destacar los siguientes puntos:

La facilidad del modelo MARS para trabajar con interacción entre variables hacen que este método sea muy atractivo pues como se evidenció en la sección 3.1 se puede encontrar como mejor modelo, un modelo aditivo (o modelo lineal para el año 2020) o un modelo con interacción de grado 3 (año 2019) sin tener la necesidad de cambiar de técnica de modelado.

Contar con una gran cantidad de cantones con muestra representativa hace que mejoren las estimaciones hechas por el modelo MARS lo que se puede observar en el poder predictivo del modelo para el año 2019.

Por otra parte, realizar este procedimiento depende tanto de la encuesta ENEMDU como medio para encontrar el ingreso promedio per cápita por cantón como de una base de datos con registros completos para todos los cantones como el valor agregado bruto a nivel cantonal, haciendo de nuevo hincapié en la importancia de contar con información a este nivel de desagregación.

Un mapa de calor es una herramienta exploratoria que sirve de guía para identificar a priori como están distribuidas las unidades espaciales y de la existencia de algún patrón de asociación, lo que se puede corroborar mediante la aplicación de un índice de autocorrelación espacial. Particularizando a nuestro caso de estudio, una vez que se tiene el ingre-

so promedio per cápita para todos los cantones, el análisis de autocorrelación espacial evidenció la existencia de este para los años analizados; para todos ellos se considera autocorrelación espacial positiva, aunque no es muy fuerte para 2019 y 2021.

Los efectos de la pandemia COVID-19 se pueden observar en la evolución de los mapas de asociación espacial para los años en estudio (ver Figura 3.10). A pesar de que la provincia del Oro se mostraba como una de las de mayor crecimiento económico, con algunos cantones de ella las que forman el primer clúster Alto-Alto, al inicio de la pandemia en el año 2020 fue una de las más afectadas siendo la segunda provincia con la mayor cantidad de fallecidos [12]; demás, el sector bananero (pilar económico de la provincia) tuvo un decremento económico al disminuir su producción en 2,29% con respecto al año anterior [11] lo que incurre en aumento de costos y gastos para tratar de mantener la misma producción.

Sin embargo, el sector Amazónico se destaca al presentar el mayor clúster con la característica Bajo-Bajo, que, entre otros factores, el coronavirus afecto a la economía de la mayoría de los pueblos residentes en esta zona del país, pues significó un gran reto en temas de transporte, educación, salud, incurriendo en grandes costos para los ciudadanos debido a las condiciones geográficas de la región. El análisis de autocorrelación espacial local se muestra como una herramienta para conocer la dinámica de asociación de las unidades espaciales, siendo de utilidad para enfocar nuestros estudios en zonas específicas del espacio.

Se recomienda utilizar información actualizada como el censo que se lleva a cabo a finales de 2022 y realizar una comparación entre la metodología propuesta y un método de áreas pequeñas.

Debido a la existencia de autocorrelación espacial por medio del índice I de Moran así como la formación de clúster espaciales, se recomienda para futuros trabajos realizar estimaciones del ingreso promedio per cápita por cantones en el Ecuador mediante modelos que consideren efectos espaciales.

Se recomienda estudiar con mayor profundidad la realidad económica de los cantones que muestran tener ingresos bajos y que sus vecinos

tienen ingresos altos o el caso contrario para la correcta generación de políticas que ayuden a su crecimiento económico.

Se recomienda explorar más metodologías en el cálculo del ingreso promedio per cápita que tomen en cuenta este nivel de desagregación, puesto que los cantones son el segundo nivel en la división política que tiene Ecuador y así poder expandir la cantidad de información para este nivel agregado.

Capítulo A

Anexos

A.1. Codificación de provincias y cantones de Ecuador

Código Provincia	Provincia
01	AZUAY
02	BOLÍVAR
03	CAÑAR
04	CARCHI
05	COTOPAXI
06	CHIMBORAZO
07	EL ORO
08	ESMERALDAS
09	GUAYAS
10	IMBABURA
11	LOJA
12	LOS RÍOS
13	MANABÍ
14	MORONA SANTIAGO
15	NAPO
16	PASTAZA
17	PICHINCHA

18	TUNGURAHUA
19	ZAMORA CHINCHIPE
20	GALÁPAGOS
21	SUCUMBÍOS
22	ORELLANA
23	SANTO DOMINGO DE LOS TSÁCHILAS
24	SANTA ELENA

Los dos primeros dígitos del código cantón hacen referencia a su provincia y los dos últimos dígitos lo identifican dentro de ella.

Código Cantón	Cantón	Código Cantón	Cantón
0101	CUENCA	1106	ESPINDOLA
0102	GIRON	1107	GONZANAMA
0103	GUALACEO	1108	MACARA
0104	NABON	1109	PALTAS
0105	PAUTE	1110	PUYANGO
0106	PUCARA	1111	SARAGURO
0107	SAN FERNANDO	1112	SOZORANGA
0108	SANTA ISABEL	1113	ZAPOTILLO
0109	SIGSIG	1114	PINDAL
0110	OÑA	1115	QUILANGA
0111	CHORDELEG	1116	OLMEDO LOJA
0112	EL PAN	1201	BABAHOYO
0113	SEVILLA DE ORO	1202	BABA
0114	GUACHAPALA	1203	MONTALVO
0115	CAMILO PONCE ENRIQUEZ	1204	PUEBLOVIEJO
0201	GUARANDA	1205	QUEVEDO
0202	CHILLANES	1206	URDANETA

0203	CHIMBO	1207	VENTANAS
0204	ECHEANDIA	1208	VINCES
0205	SAN MIGUEL DE BOLIVAR	1209	PALENQUE
0206	CALUMA	1210	BUENA FE
0207	LAS NAVES	1211	VALENCIA
0301	AZOGUES	1212	MOCACHE
0302	BIBLIAN	1213	QUINSALOMA
0303	CAÑAR	1301	PORTOVIEJO
0304	LA TRONCAL	1302	BOLIVAR CALCETA
0305	EL TAMBO	1303	CHONE
0306	DELEG	1304	EL CARMEN
0307	SUSCAL	1305	FLAVIO ALFA- RO
0401	TULCAN	1306	JIPIJAPA
0402	BOLIVAR (CARCHI)	1307	JUNIN
0403	ESPEJO	1308	MANTA
0404	MIRA	1309	MONTECRISTI
0405	MONTUFAR	1310	PAJAN
0406	SAN PEDRO DE HUACA	1311	PICHINCHA
0501	LATACUNGA	1312	ROCAFUERTE
0502	LA MANA	1313	SANTA ANA
0503	PANGUA	1314	SUCRE
0504	PUJILI	1315	TOSAGUA
0505	SALCEDO	1316	24 DE MAYO
0506	SAQUISILI	1317	PEDERNALES
0507	SIGCHOS	1318	OLMEDO MA- NABI
0601	RIOBAMBA	1319	PUERTO LO- PEZ
0602	ALAUSI	1320	JAMA
0603	COLTA	1321	JARAMIJO
0604	CHAMBO	1322	SAN VICENTE

0605	CHUNCHI	1401	MORONA
0606	GUAMOTE	1402	GUALAQUIZA
0607	GUANO	1403	LIMON IN- DANZA
0608	PALLATANGA	1404	PALORA
0609	PENIPE	1405	SANTIAGO
0610	CUMANDA	1406	SUCUA
0701	MACHALA	1407	HUAMBOYA
0702	ARENILLAS	1408	SAN JUAN BOSCO
0703	ATAHUALPA	1409	TAISHA
0704	BALSAS	1410	LOGROÑO
0705	CHILLA	1411	PABLO SEX- TO
0706	EL GUABO	1412	TIWINTZA
0707	HUAQUILLAS	1501	TENA
0708	MARCABELI	1502	ARCHIDONA
0709	PASAJE	1503	EL CHACO
0710	PIÑAS	1504	QUIJOS
0711	PORTOVELO	1505	CARLOS JULIO ARO- SEMENA TOLA
0712	SANTA ROSA	1601	PASTAZA
0713	ZARUMA	1602	MERA
0714	LAS LAJAS	1603	SANTA CLARA
0801	ESMERALDAS	1604	ARAJUNO
0802	ELOY ALFARO	1701	QUITO
0803	MUISNE	1702	CAYAMBE
0804	QUININDE	1703	MEJIA
0805	SAN LOREN- ZO	1704	PEDRO MON- CAYO
0806	ATACAMES	1705	RUMIÑAHUI
0807	RIOVERDE	1707	SAN MIGUEL DE LOS BAN- COS

0901	GUAYAQUIL	1708	PEDRO VICENTE MALDONADO
0902	ALFREDO BAQUERIZO MORENO	1709	PUERTO QUITO
0903	BALAO	1801	AMBATO
0904	BALZAR	1802	BAÑOS
0905	COLIMES	1803	CEVALLOS
0906	DAULE	1804	MOCHA
0907	DURAN	1805	PATATE
0908	EMPALME	1806	QUERO
0909	EL TRIUNFO	1807	PELILEO
0910	MILAGRO	1808	PILLARO
0911	NARANJAL	1809	TISALEO
0912	NARANJITO	1901	ZAMORA
0913	PALESTINA	1902	CHINCHIPE
0914	PEDRO CARBO	1903	NANGARITZA
0916	SAMBORONDONI	1904	YACUAMBI
0918	SANTA LUCIA	1905	YANZATZA
0919	SALITRE (URBINA JADO)	1906	EL PANGUI
0920	YAGUACHI	1907	CENTINELA DEL CONDOR
0921	PLAYAS	1908	PALANDA
0922	SIMON BOLIVAR	1909	PAQUISHA
0923	CORONEL MARCELINO MARIDUEÑA	2101	LAGO AGRIO
0924	LOMAS DE SARGENTILLO	2102	GONZALO PIZARRO
0925	NOBOL	2103	PUTUMAYO

0927	GENERAL ANTONIO ELIZALDE	2104	SHUSHUFINDI
0928	ISIDRO AYO- RA	2105	SUCUMBIOS
1001	IBARRA	2106	CASCALES
1002	ANTONIO AN- TE	2107	CUYABENO
1003	COTACACHI	2201	ORELLANA
1004	OTAVALO	2202	AGUARICO
1005	PIMAMPIRO	2203	LA JOYA DE LOS SACHAS
1006	SAN MIGUEL DE URCUQUI	2204	LORETO
1101	LOJA	2301	SANTO DO- MINGO
1102	CALVAS	2302	LA CONCOR- DIA
1103	CATAMAYO	2401	SANTA ELENA
1104	CELICA	2402	LA LIBERTAD
1105	CHAGUARPAMBA	2403	SALINAS

A.2. Cantones con muestra representativa para los años 2020 y 2021

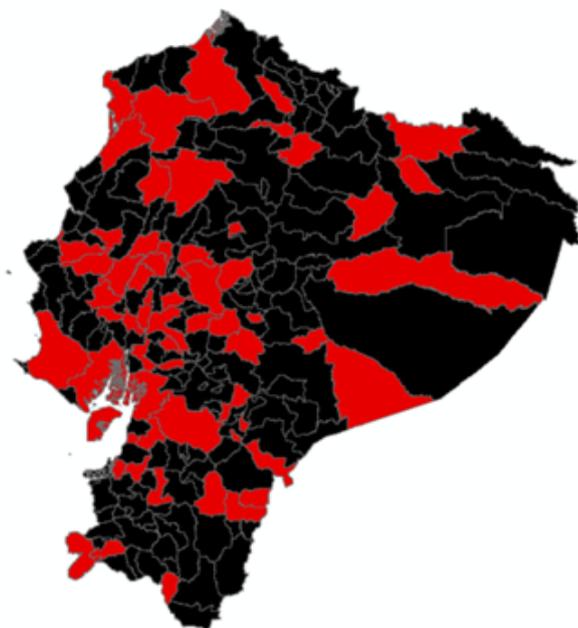


Figura A.1: Cantones con muestra representativa año 2020.

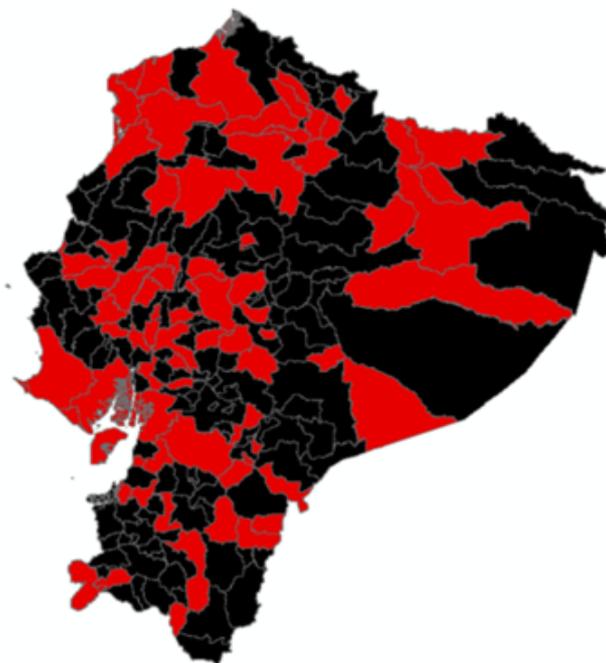


Figura A.2: Cantones con muestra representativa año 2021.

A.3. Importancia de las variables del modelo MARS y gráficos de dispersión para muestra de entrenamiento y validación para el año 2020.

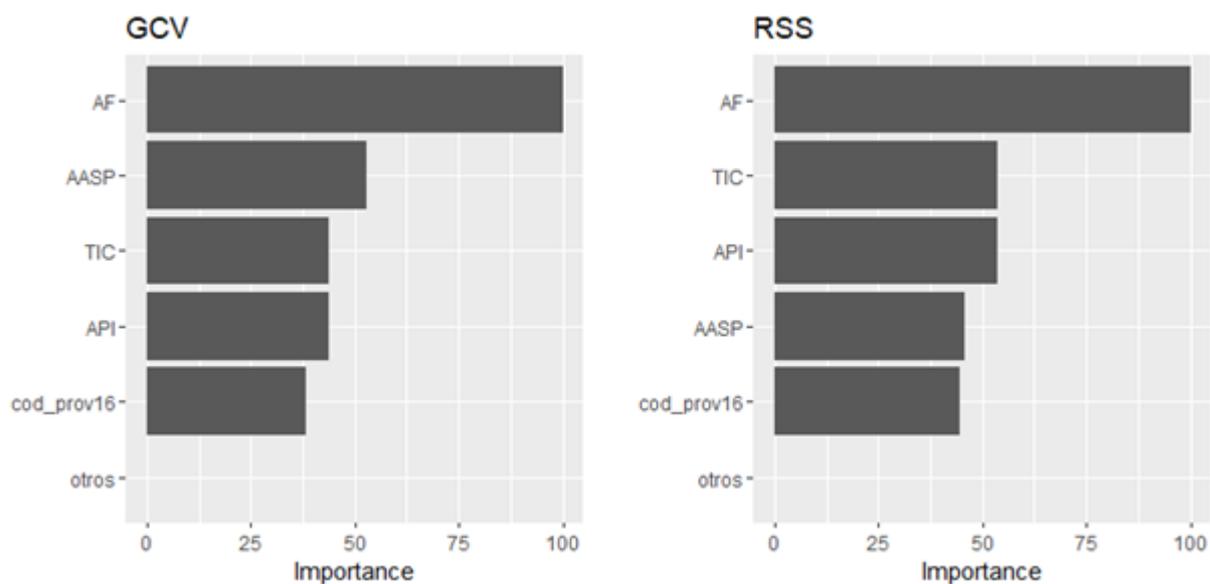


Figura A.3: Importancia de las variables basado en el impacto de los valores de GCV y RSS para el año 2020.

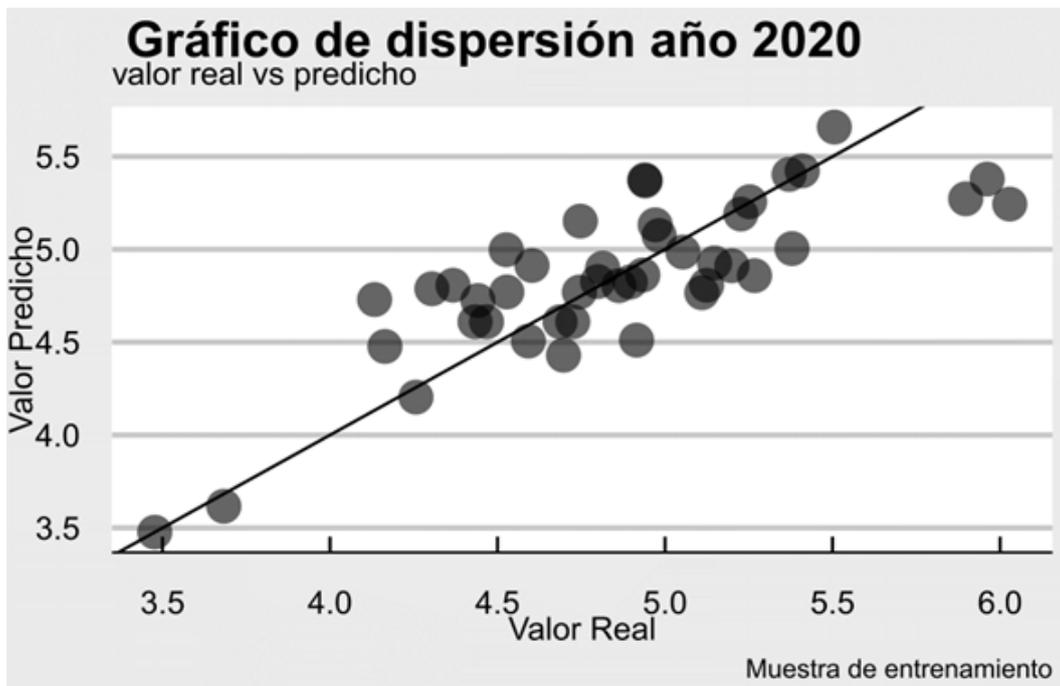


Figura A.4: Valores reales y predicho para la muestra de entrenamiento (train) para el año 2020.

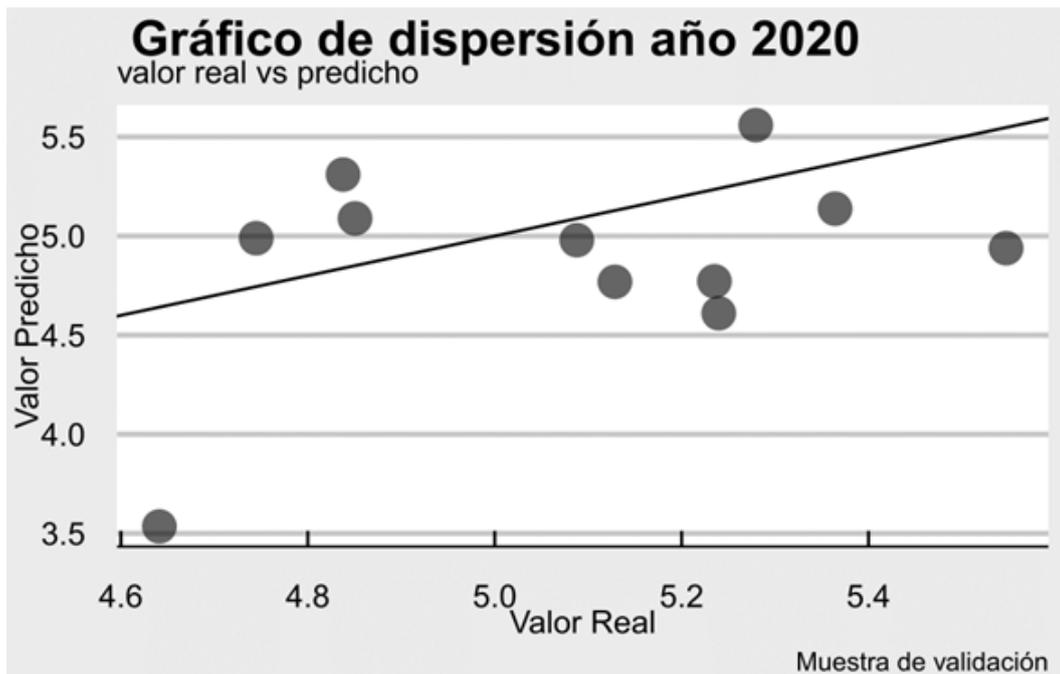
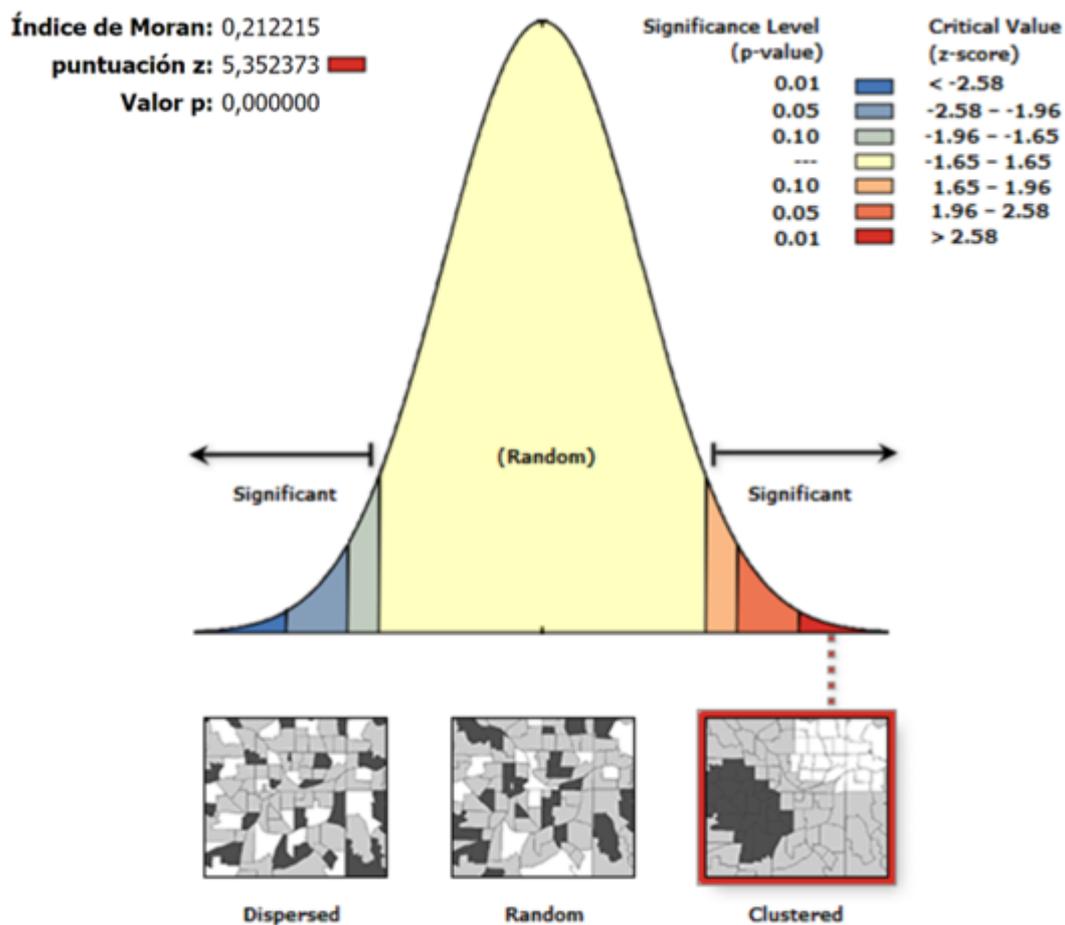


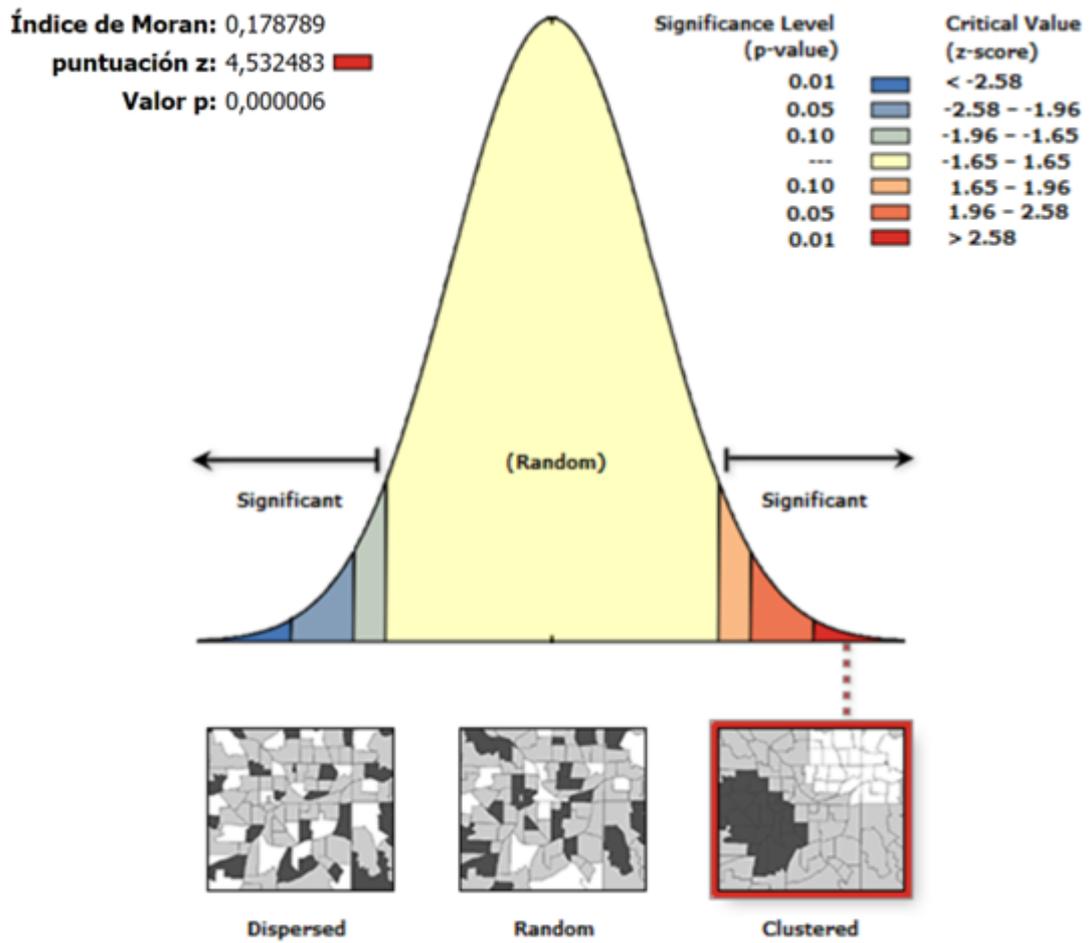
Figura A.5: Valores reales y predicho para la muestra de entrenamiento (test) para el año 2020.

A.4. Informes de autocorrelación espacial y mapas de agrupamiento clúster para los años 2020 y 2021.



Dada la puntuación z de 5.35237303831, existe una probabilidad menor al 1% de que el patrón agrupado pueda ser el resultado de una verosimilitud aleatoria.

Figura A.6: Informe de Autocorrelación Espacial para el año 2020



Dada la puntuación z de 4.53248333801, existe una probabilidad menor al 1% de que el patrón agrupado pueda ser el resultado de una verosimilitud aleatoria.

Figura A.7: Informe de Autocorrelación Espacial para el año 2021

Referencias bibliográficas

- [1] Luc Anselin. Local indicators of spatial association—lisa. *Geographical analysis*, 27(2):93–115, 1995.
- [2] Mercy Orellana Bravo, Rodrigo Mendieta Muñoz, and Nelson Tapia Olvera. La pobreza multidimensional en el ecuador. un análisis espacial.
- [3] Álvaro Briz-Redón, Jorge Mateu, and Francisco Montes. Identifying crime generators and spatially overlapping high-risk areas through a nonlinear model: A comparison between three cities of the valencian region (spain). *Statistica Neerlandica*, 76(1):97–120, 2022.
- [4] Grace Estefanía Hidalgo Bucheli. Uso del índice de moran y lisa para explicar el ausentismo electoral rural en ecuador. *Revista Geográfica*, (160):91–108, 2019.
- [5] Wilton O Bussab and M Ferreira. Critério brasil de estratificação socioeconômica: Aspectos demográficos. *Proceedings of CLADEA, San Juan, Porto Rico*, 34, 1999.
- [6] Davis Xavier Salinas Campuzano, Joselyn Ivonne Cedeño Alonso, Flor Yelena Vega Jaramillo, and Jorge Guido Sotomayor Pereira. El valor agregado bruto del ecuador 2007-2017: The gross value added of ecuador 2007-2017. In *Conference Proceedings UTMACH*, volume 5, pages 9–27, 2021.
- [7] Alejandro Cañadas. Does spatial clustering help explaining differences in the inequality of income distribution?: evidence from argentina. 2010.

- [8] Rubén Fernández Casal and Julián Costa. Aprendizaje estadístico, 2020.
- [9] Juan Pablo Celemín. Autocorrelación espacial e indicadores locales de asociación espacial: Importancia, estructura y aplicación. *Revista Universitaria de Geografía*, 18(1):11–31, 2009.
- [10] NU CEPAL et al. El impacto del covid-19 en los pueblos indígenas de américa latina-abya yala: entre la invisibilización y la resistencia colectiva. 2020.
- [11] Melany Maybrith Loja Chillogallo, Milena Maylen Saavedra Jaramillo, and Laura Luzmila Vega González. Covid-19: Impacto económico al sector bananero de la provincia de el oro, ecuador, período 2019-2020. *Polo del Conocimiento: Revista científico-profesional*, 7(2):24, 2022.
- [12] EL COMERCIO. El oro es la segunda provincia con más muertes por covid-19. actualidad-el comercio. <https://www.elcomercio.com/actualidad/ecuador/oro-muertes-covid-coronavirus-ecuador.html>, Abril 2020.
- [13] EL COMERCIO. Santo domingo recuperó el comercio; los contagios de covid-19 suben. actualidad-el comercio. <https://www.elcomercio.com/actualidad/ecuador/santo-domingo-comercio-contagios-covid19.html>, Septiembre 2020.
- [14] Noel Cressie and John Kornak. Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical science*, pages 436–456, 2003.
- [15] Instituto Nacional de Estadística y Censos. Encuesta nacional de empleo, desempleo y subempleo (enemdu), junio 2021. https://www.ecuadorencifras.gob.ec/documentos/web-inec/POBREZA/2021/Junio-2021/202106_Boletin_pobreza.PDF, Julio 2021.

- [16] Secretaría Nacional de Planificación (s.f.). Proyecciones y estudios demográficos. <https://sni.gob.ec/proyecciones-y-estudios-demograficos>, 2021.
- [17] Banco Central del Ecuador (s.f.). Cuentas nacionales regionales. <https://contenido.bce.fin.ec/documentos/Estadisticas/SectorReal/CuentasCantonaes/Indice.htm>, 2022.
- [18] Roberto Ezcurra, Pedro Pascual, and Manuel Rapún. The spatial distribution of income inequality in the european union. *Environment and Planning A*, 39(4):869–890, 2007.
- [19] Marie-Josée Fortin, Pierre Drapeau, and Pierre Legendre. Spatial autocorrelation and sampling design in plant ecology. In *Progress in theoretical vegetation science*, pages 209–222. Springer, 1990.
- [20] Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67, 1991.
- [21] S. Milborrow. Derived from mda:mars by T. Hastie and R. Tibshirani. earth: Multivariate adaptive regression splines . r package. 2011.
- [22] Robert C Geary. The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146, 1954.
- [23] A Getis, J Ord, L Anselin, and SJ Rey. The analysis of spatial association by use of distance statistics. advances in spatial science. In *Perspectives on spatial data analysis*, pages 127–145. Springer Berlin, 2010.
- [24] María Pilar González Casimiro. Técnicas de predicción económica. 2009.
- [25] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *OTM Confederated International Conferences. on the Move to Meaningful Internet Systems*, pages 986–996. Springer, 2003.
- [26] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

- [27] Karen Kemp et al. *Encyclopedia of geographic information science*. Sage, 2008.
- [28] Fernando Medina. *Consideraciones sobre el índice de Gini para medir la concentración del ingreso*. Cepal, 2001.
- [29] Andrea Molina, E Cabrera, L Moreno, MA Sharman, and F Cuevas. Mapa de pobreza y desigualdad por consumo ecuador 2014. *Instituto Nacional de Estadística y Censos y Banco Mundial (INEC-BM)*. Quito-Ecuador, 2015.
- [30] Patrick AP Moran. The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251, 1948.
- [31] Banco Mundial. Índice de gini - ecuador. <https://datos.bancomundial.org/indicador/SI.POV.GINI?locations=EC>, Mayo 2020.
- [32] José Ollague, Lenny Capa, Ernesto Novillo, Tatiana Sanchez, Liana Sánchez, and María García. Variables sociales, económicas y productivas como referente de posicionamiento nacional de la provincia de el oro, ecuador. *Revista Espacios*, 40(37):13, 2019.
- [33] PRIMICIAS. Las razones por las que esmeraldas se desangra. <https://www.primicias.ec/noticias/firmas/esmeraldas-violencia-pobreza-ecuador/>, Septiembre 2022.
- [34] Liliana Ramírez and Vilma Falcón. Autocorrelación espacial: Analogías y diferencias entre el índice de moran y el índice getis y ord. aplicaciones con indicadores de acceso al agua en el norte argentino. *Ponencia presentada en las Jornadas Argentinas de Geotecnologías, Universidad Nacional de San Luis*, 2, 2015.
- [35] J.N. Rao and I. Molina. *Small Area Estimation*. 2015 <https://doi.org/10.1002/9781118735855>.
- [36] Damián Rivadeneira, Daniela Sandoval, Diana Zambonino, and Andrés Albán. Encuesta nacional de empleo, desempleo y subempleo (enemdu) documento metodológico. *Quito: Instituto Nacional de Estadística y Censos*, 2018.

- [37] Manuel Ruiz, Fernando López, and Antonio Páez. Testing for spatial association of qualitative data using symbolic dynamics. *Journal of Geographical Systems*, 12(3):281–309, 2010.
- [38] Victoria Salinas Castro, Richard E Bilborrow, and Clark Gray. Cambios socioeconómicos en el siglo xxi en poblaciones indígenas amazónicas: retos actuales. *Estudios demográficos y urbanos*, 35(1):83–116, 2020.
- [39] Rosina Moreno Serrano and Esther Valcarce Vayá. Econometría espacial: nuevas técnicas para el análisis regional. una aplicación a las regiones europeas. *Investigaciones Regionales-Journal of Regional Research*, (1):83–106, 2002.
- [40] Willington Siabato and Jhon Guzmán-Manrique. Spatial autocorrelation and the development of quantitative geography. *Cuadernos de Geografía: Revista Colombiana de Geografía*, 28(1):1–22, 2019.
- [41] Justin J Slater, Patrick E Brown, Jeffrey S Rosenthal, and Jorge Mateu. Capturing spatial dependence of covid-19 case counts with cellphone mobility data. *Spatial Statistics*, 49:100540, 2022.
- [42] Ivette Rossana Vallejo and Katy Alvarez. La pandemia del coronavirus en la amazonía ecuatoriana: vulnerabilidades y olvido del estado. *Cadernos de Campo (São Paulo-1991)*, 29(1):94–110, 2020.
- [43] Jairo Vanegas and Fabián Vásquez. Multivariate adaptative regression splines (mars), una alternativa para el análisis de series de tiempo. *Gaceta sanitaria*, 31:235–237, 2017.