

# EL USO DE ÁRBOLES DE DECISIÓN COMO HERRAMIENTA PARA GENERAR UN MODELO DE VALORACIÓN TERRITORIAL EN VILCABAMBA, ECUADOR.

Fabián Reyes-Bueno

Grupo de Investigación 1934-TB-Laboratorio de Territorio (Dep. Ingeniería Agroforestal-IBADER)  
Universidad de Santiago de Compostela. Escuela Politécnica Superior de Lugo, España  
f.reyes@usc.es

David Miranda-Barrós

Grupo de Investigación 1934-TB-Laboratorio de Territorio (Dep. Ingeniería Agroforestal-IBADER)  
Universidad de Santiago de Compostela. Escuela Politécnica Superior de Lugo, España  
david.miranda@usc.es

Rafael Crecente-Maseda

Grupo de Investigación 1934-TB-Laboratorio de Territorio (Dep. Ingeniería Agroforestal-IBADER)  
Universidad de Santiago de Compostela. Escuela Politécnica Superior de Lugo, España  
rafael.crecente@usc.es

**RESUMEN:** El valor de la tierra es una variable difícil de determinar, debido no solamente a factores no visibles pero presentes en el proceso de transacción, sino también a las características locales que influyen sobre la formación del precio. Las particularidades a nivel local no siempre afectan de la misma manera al valor de las tierras, por lo que se hace necesario aplicar técnicas que se adapten adecuadamente a las relaciones entre factores o variables y el valor. En este trabajo se aplicó el algoritmo M5P para modelar el valor de la tierra. Para ello se contó con una muestra de 132 transacciones obtenidas entre junio de 2008 a julio de 2009 en el área rural de la parroquia Vilcabamba, Ecuador. Los resultados reflejan que con la aplicación de Árboles de modelo se logra estimar de forma más precisa del valor en comparación con regresión lineal, generando un modelo para cada grupo de muestras con comportamientos similares, se mejora la capacidad explicativa del comportamiento del valor en el territorio, evidenciando la influencia de la localización en el establecimiento de este valor.

**Palabras clave:** valoración de tierras, árboles de decisión, Vilcabamba, SIG.

## 1. INTRODUCCIÓN

La valoración de la tierra es una actividad ampliamente extendida desde hace muchos siglos atrás. La razón fundamental es que ésta fue y es la principal fuente de riqueza de los pueblos. Como lo menciona Buurman (2003), en los siglos XVII y XVIII la tierra fue considerada un importante factor de producción, pero teorías económicas modernas tratan a la tierra además como un bien de consumo o como un bien de inversión.

Como principal fuente de riqueza, la tierra ha sido el elemento principal para la implantación de impuestos.

Por otro lado, el catastro se puede definir como “un sistema de información de tierras actualizado y basado en la parcela, conteniendo un registro de intereses en la tierra” (Williamson 2001).

Por lo tanto, podríamos definir la valoración catastral como la estimación del valor de la tierra basada en la parcela (o la unidad territorial). Generalmente sigue un procedimiento reglado, y es una valoración masiva.

Como lo menciona Stumpf (2007) para fines tributarios, el principal requisito es la garantía de equidad, es decir, que el nivel de precisión sea similar para todos los inmuebles. Esos requisitos deben ser objetivos, precisos y explicables.

El criterio frecuentemente aplicado para la valoración catastral rural es la capitalización de la renta, en la que es asignado un valor al inmueble en función de la renta que puede fijar en un determinado período de tiempo, pero ¿es actualmente el criterio aplicable de acuerdo a la importancia hacia la que ha evolucionado la tierra? ¿es aplicable en territorios que pueden tener amplias zonas con una renta nula?

Una alternativa a ello es considerar el criterio de valor de mercado como una forma de brindar un mayor acercamiento a los factores (de productividad o no) que influyen en el valor del predio.

Uno de los métodos econométricos más utilizados, para generar modelos de valoración, es el de regresión lineal, pero su aplicabilidad en este campo en los últimos tiempos ha sido ampliamente discutida por no ser el método adecuado para modelar las relaciones no lineales existentes entre muchas variables y el valor de la tierra.

Hoy en día, la mejora en la tecnología computacional, el aumento de la cantidad y calidad de información territorial y el nacimiento de software con interfaz amigable, permiten aplicar otras técnicas posiblemente más apropiadas para obtener variables espaciales (utilizando SIG) para modelar el territorio, y técnicas de Valoración Automatizada AVM.

Una de las principales ventajas que presenta es la facilidad de recuperación y procesamiento de grandes volúmenes de información. La precisión del modelo de valoración generado depende en gran medida de tres aspectos: la calidad de la información, la cantidad suficiente de datos que ofrezcan representatividad, y el método utilizado para generar el modelo.

Existen alternativas que entran en el campo de la minería de datos, tales como los modelos lineales generalizados, redes neuronales artificiales, métodos difusos, entre otros. Kauko y d'Amato (2009) citan entre otras, el análisis multinivel (Orford, 2002), coeficientes de variación espacial, modelos aditivos generalizados, algoritmos genéticos, sistemas expertos basados en reglas, lógica difusa, regresión con peso geográfico o la integración de técnicas de inteligencia artificial...

La segmentación del mercado es una alternativa para modelar el valor de la tierra. Kennedy et al. (1997) utilizan análisis de componentes principales y técnicas de análisis de grupos para identificar submercados rurales. Hayles (2006) comparó los resultados del establecimiento de regresión en submercados segmentados geográficamente versus submercados generados con técnicas de análisis de grupos (a-priori) para modelar el precio de la tierra, obteniendo mejores resultados con la segunda alternativa.

Una alternativa para el modelado de segmentos de mercado son los árboles de modelo, los cuales toman la forma de árboles de decisión y plantean funciones de regresión en vez de clases para cada hoja. La técnica fue aplicada por Acciani et al. (2008) para modelar el precio de 109 propiedades de viñedos en el Sur de Italia, logrando mejores resultados que MRA y un modelo comprensible.

Una de las limitaciones al considerar la utilización de técnicas basadas en árboles de modelo es la inestabilidad ante variaciones de la muestra.

El catastro y la valoración catastral en Ecuador, es una tarea que a partir del año 1998, a través del proceso de descentralización, fue encomendada a los municipios del país. De acuerdo a la Ley de Régimen Municipal (2005), se establece que el concejo municipal aprobará los factores de aumento o reducción del valor (Art. 332), y que el valor del suelo será determinado por un proceso de comparación con precios de venta de parcelas de condiciones similares del mismo sector (307).

Debido a que los propios técnicos municipales no han tenido ningún tipo de formación para llevar adelante estas tareas, en una gran parte del territorio el método utilizado para valorar la tierra podría estar muy lejos del valor real de los predios.

El objetivo del presente estudio es elaborar un modelo de valoración de tierras adaptado a las características no lineales de las variables que lo componen, y compararlo con el método tradicional de regresión lineal.

## 2. METODOLOGÍA Y DATOS

El área de estudio se centra en la parroquia Vilcabamba, ubicada en la provincia de Loja, al sur del Ecuador.



**Figura 1. Ubicación espacial de la parroquia Vilcabamba.**

La investigación de campo tuvo lugar entre junio de 2008 y julio de 2009, contando al final con un número total de 132 transacciones de tierras y diversas características de calidad de acceso al predio, localización, servicios básicos o características estructurales, obtenidos para un período de 6 años.

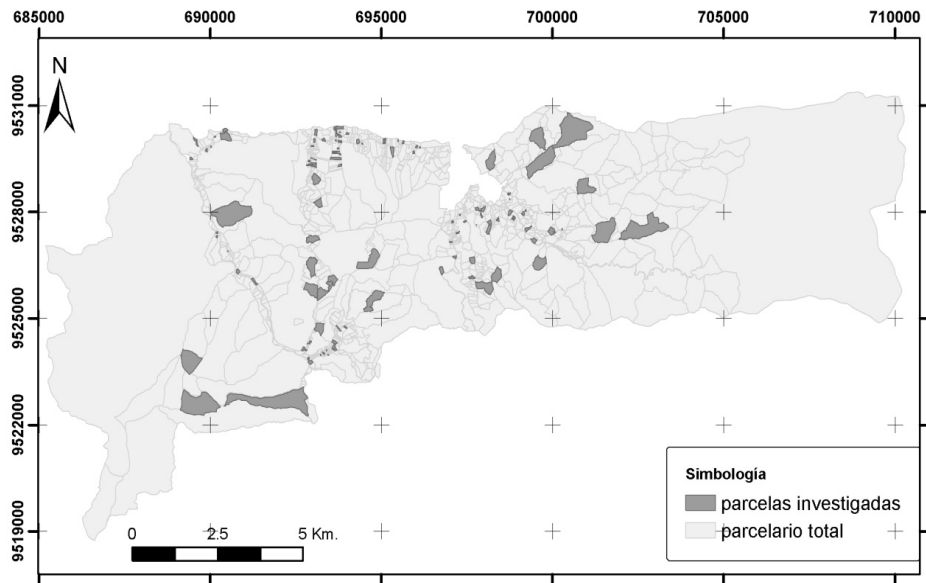


Figura 2. Distribución espacial de las parcelas investigadas en la parroquia Vilcabamba.

En la Tabla 1 y 2 se describen las variables continuas y discretas utilizadas para modelar el valor de la tierra.

Tabla 1. Tabla resumen de las variables continuas recopiladas para el análisis.

Nombre de la variable	Descripción	Nombre de la variable	Descripción
SUPERFICIE	Superficie en m <sup>2</sup>	dist_t_car	Distancia a carretera en metros
t_abs_ce	Tiempo real al centro económico en minutos	i_d_ce	Índice de distancia al centro económico
D_ABS_CPOB	Distancia real al centro poblado en metros	CULTIVO	Proporción de cultivo
CP	Proporción de cultivo permanente	PASTO	Proporción de pasto
BOSQUE	Proporción de bosque	MATO	Proporción de matorral
HABITA	Proporción de área edificada	RIEGO	Proporción de riego

Tabla 2. Tabla resumen de las variables discretas recopiladas para el análisis.

VARIABLE	Descripción	PARÁMETRO	VARIABLE	Descripción	PARÁMETRO
FIRME	Firme de la vía	LA: Lastrado RA: Pavimentado, TI: Tierra	ESTADO	Estado de la vía	B: Bueno, M: Malo, R: Regular
COBERTURA	Cobertura del suelo	BOSQUE, CP: Cultivo Permanente, CULTIVOS, MATORRAL, PASTO	PENDIENTE	Tipo de pendiente	PEND1: Débil, PEND2: Suave, PEND3: Moderada, PEND4: Fuerte, PEND5: Muy Fuerte, PEND6: Abrupta
AGUA	Conexión a agua	C: Conectado, NC: No conectado	LUZ	Conexión a energía eléctrica	C: Conectado, NC: No conectado

Fueron transformadas algunas variables continuas con su Logaritmo Natural: superficie, tiempo de acceso al centro económico, valor actual, distancia hacia centros poblados y distancia desde la parcela a la carretera. En el caso de la distancia tanto hacia centros poblados como a carretera (por existir valores con valores de cero), en primer lugar se ha incrementado unidad para luego poder calcular su logaritmo.

Adicionalmente fue generada una variable de valor influenciada espacialmente, la cual guarda un precio medio ponderado de cada predio, formado por la ponderación de los valores de las parcelas vecinas, obtenida en base a una matriz de pesos espaciales en función de la distancia inversa a través de la red de acceso a las parcelas. Para comparar los modelos, las variables fueron normalizadas.

El proceso de análisis fue desarrollado utilizando el software Weka (Hall et al., 2009). Fueron evaluadas dos bases de datos diferentes (una de ellas eliminando la variable espacial 'red\_ponderado') con dos algoritmos (árbol de modelos y regresión lineal).

El árbol de modelo se ha generado tomando como base el algoritmo M5P, el cual tiene las siguientes características:

- Construye segmentos (nodos) basados en el tratamiento de la desviación estándar de los valores de clase de la base de datos, minimizando la variación de los elementos en los valores de clase agrupados bajo cada nuevo nodo generado.
- Para prevenir un sobreajuste por la generación de un árbol con una gran cantidad de hojas que no sea capaz de modelar el comportamiento cuando sea aplicado a otro subconjunto de datos de la población, es conveniente obtener un modelo más general, eliminando condiciones demasiado específicas, proceso conocido como poda.

Una descripción más detallada del algoritmo puede consultarse en Witten y Frank (2005).

Para el análisis de regresión múltiple primeramente se utilizó el mismo proceso de selección de variables del algoritmo M5P, con el cual se diseñó el modelo teniendo en cuenta únicamente las variables más relevantes.

La bondad de los modelos fue analizada atendiendo a los siguientes evaluadores (Witten y Frank, 2005):

- Coeficiente de correlación CC (que mide la correlación estadística entre los valores estimados y los actuales del valor)
- Raíz del error medio cuadrático RMSE (está medido en la misma unidad de la variable dependiente. La situación ideal sería cuando este error sea igual a 0)
- Error absoluto medio MAE (tiene únicamente en cuenta la magnitud de los errores individuales sin tener en cuenta su signo)
- Raíz del error cuadrático relativo RRSE (calcula el error relativo con respecto al promedio de los valores actuales de la variable dependiente)
- Error relativo de los valores absolutos RAE (es el error absoluto medio que tiene el mismo tipo de normalización que el RRSE).

Se analizó el grado de influencia existente en cada una de las variables finalmente utilizadas en el modelo y el valor.

El modelo además fue evaluado calculando el coeficiente de dispersión (COD), que indica la desviación media de un grupo del radio desde la mediana expresado como un porcentaje. De acuerdo a los estándares establecidos por la IAAO (2010), en el área rural el COD aceptable varía entre 5 a 30%. En sectores ubicados en el área rural, dedicados a residencia, se acepta un COD de 20%, en aquellos predios destinados a la producción se acepta un COD de hasta 25% y en aquellos predios rurales de tierra vacante no destinada a la agricultura se acepta un COD de hasta 30%.

Adicionalmente fue calculado el error relativo de cada observación para determinar la frecuencia existente al 5, 10, 15 y 20% y mayor del 50%.

### 3. RESULTADOS EMPÍRICOS

Los modelos generados a partir de regresión lineal fueron construidos con 11 y 12 variables dependiendo de la base de datos utilizada. Se observa una fuerte participación de la superficie en el valor de la tierra, pero también la accesibilidad juega un papel importante en este valor.

**Tabla 3. Modelo de valoración de tierras obtenido a través de regresión lineal.**

Variable	BBDD1	BBDD2
ln_sup	-3.1462	-3.0697
FIRME: LA	0.1956	0.2625
FIRME: RA	0.1926	0.2625
ln_t_ce	-1.7959	-1.4986
ln_d_c_pob_1	-0.4065	-0.3448
i_d_ce	1.2803	0.9751
PEND4	0.3668	0.3735
PEND3	0.3668	0.3735
PEND2	0.708	0.7055
PEND1	0.3668	0.7055
AGUA	0.2747	0.3819
red_ponderado		1.2903
Intercección	1.3703	0.8679
R2	0.836	0.841
R2 ajustado	0.821	0.824

La distancia existente hacia centros poblados, la superficie y el valor ponderado son las variables en función de las que se generaron segmentos de mercado.

La participación de la superficie varía dependiendo del segmento, siendo muy influyente sobre todo en parcelas ubicadas a más de 460 metros de centros poblados o con superficies mayores a 4800 metros cuadrados.

Se observa una influencia positiva si la vía de acceso es pavimentada o lastrada, siendo más valorada sobre todo cuando la parcela se aleja de los centros poblados.

El tiempo de acceso también ejerce una fuerte influencia sobre todo sobre aquellos predios con pequeña superficie.

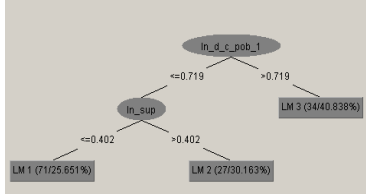
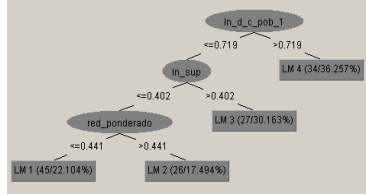
La distancia de acceso al centro económico no presenta una fuerte influencia, debido principalmente a que es la variable según la cual se segmenta la muestra. Se observa también

con respecto a esta variable, que su influencia es reemplazada por el valor ponderado cuando éste está presente.

La calidad de acceso (determinada por el índice de acceso al centro económico) ejerce más influencia sobre aquellos predios de grandes superficies y más alejados del centro poblado.

La pendiente débil, suave o moderada (predios con pendiente  $\leq 25\%$ ) ejerce un impacto positivo sobre el valor.

**Tabla 4. Modelo de Valoración de tierras obtenido a través del algoritmo M5P.**

Variable	BBDD1			BBDD2			
	LM1	LM2	LM3	LM1	LM2	LM3	LM4
							
ln_sup	-0.7513	-3.4673	-4.7811	-0.7349	-0.7349	-3.4475	-4.8814
FIRME:LA	0.0258	0.0258	0.0595	0.0355	0.0355	0.0355	0.0818
FIRME:RA	0.0258	0.0258	0.0595	0.0355	0.0355	0.0355	0.0818
ln_t_ce	-1.6707	-2.1887	-0.5883	-1.5900	-0.9234	-2.1489	-0.4964
ln_d_c_pob_1	-0.0693	-0.0693	-0.1599				
i_d_ce	0.3292	1.5771	1.7819	0.2584	0.2584	1.4966	0.8705
CP	-0.1743			0.2909	0.1384	0.1058	0.0919
	0.0830	0.121	0.1076	0.2909	0.1384	0.1058	0.0919
PEND3	0.0830	0.121	0.1076	0.2909	0.1384	0.1058	0.0919
				-0.2791	0.5121	0.0668	0.1541
PEND2				0.1745	0.1745	0.1745	3.5883
PEND1	2.248	2.4817	1.5974	2.2034	1.5932	2.35	0.9325
AGUA							
red_ponderado							
Intersección							
R2	0.896			0.912			
R2ajustado	0.887			0.903			

Se desprende de la Tabla 4 que los árboles de modelo tienen una mejor capacidad de predicción, debido principalmente a que la segmentación permite mejorar la capacidad de ajuste de las muestras. Además, se tiene una capacidad interpretativa significativamente mayor con la segmentación.

La presencia del Valor ponderado (red\_ponderado), parece mejorar la capacidad explicativa del modelo, siendo incluso una variable útil para segmentar el mercado. Cuando ésta es agregada disminuye (en el caso de regresión lineal) o desaparece (en el caso de árboles de decisión) la influencia de la distancia de acceso al centro poblado.

No se observan fuertes diferencias entre las dos bases de datos analizadas utilizando el algoritmo M5P, diferencias que son más notorias cuando es diseñado el modelo por regresión y que mejora el comportamiento cuando está presente la variable 'red\_ponderado'.

Comparando los modelos por el algoritmo que los genera, existe una importante reducción del error utilizando M5P.

**Tabla 5. Resultado de los evaluadores utilizados.**

Rótulos de fila	CC	MAE	RAE	RMSE	RRSE
LRS					
bbdd1	0.87	0.67	50.88	0.82	50.07
bbdd2	0.86	0.68	50.96	0.82	50.25
M5P					
bbdd1	0.91	0.54	40.30	0.67	40.70
bbdd2	0.91	0.54	40.45	0.67	40.77

En general, evaluando los modelos generados, se observa que el coeficiente de dispersión (COD) es alto en todos los modelos, y atendiendo a los estándares establecidos por la IAAO, es superior al error aceptado. Pese a ello, cabe resaltar que el modelo generado con M5P genera un menor COD y menores errores de predicción resultados significativamente mejores que los generados por regresión lineal, y ocurre lo mismo con la adición de la variable ‘red\_ponderado’ con lo cual se logran mejores resultados que sin ella.

**Tabla 6. Frecuencia de error (expresada en porcentaje) según la base de datos.**

Rótulos de fila	M5P		RL	
	BBDD1	BBDD2	BBDD1	BBDD2
COD	48	45	53	51
0-5%ERROR	23	19	7	15
0-10%ERROR	38	37	18	25
0-15%ERROR	52	55	32	37
0-20%ERROR	64	65	42	47
>50%ERROR	25	20	31	29

#### 4. CONCLUSIONES

Las variables que tienen mayor influencia sobre el valor de mercado no ejercen esa influencia de forma continua sobre el territorio, por lo que deben aplicarse métodos alternativos que permitan modelar esas no-linealidades.

El acelerado crecimiento de la tecnología permite la utilización de herramientas que permiten modelar adecuadamente el territorio. Una de ellas son los Sistemas de Información Geográfica, que son una herramienta imprescindible para extraer información del territorio.

El valor de la tierra está fuertemente influenciado por factores de localización, accesibilidad y estructurales.

La segmentación del mercado es una estrategia útil y recomendable para valorar el territorio, y los árboles de modelo son una herramienta útil para realizar este proceso.



En árboles de decisión, la presencia de la variable con el valor ponderado influye, aunque no significativamente, en el poder predictivo del modelo generado

Presente y futuro de esta línea de investigación: En la actualidad se está analizando la utilización de una técnica de meta-aprendizaje, que permita mejorar los resultados del modelo y disminuir el sobreajuste o variación que puede presentarse en los árboles de decisión.

## 5. REFERENCIAS

**Acciani, C.; Fucilli, V. and Sardaro, R. (2008).** *Model tree: An application in real estate appraisal*. URL: <http://ideas.repec.org/p/ags/eaa109/44853.html>.

**Buurman, J. (2003).** *Rural Land Markets: a spatial explanatory model*. Amsterdam (PhD Thesis).

**CONGRESO NACIONAL DEL ECUADOR (2005).** *Codificación a la Ley Orgánica de Régimen Municipal*.

**Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. and Witten, I. (2009).** The weka data mining software: An update. *SIGKDD Explorations*, 11(1):9. URL: [http://www.kdd.org/explorations/issues/11\\_1\\_2009\\_07/p2V11n1.pdf](http://www.kdd.org/explorations/issues/11_1_2009_07/p2V11n1.pdf).

**Hayles, K. (2006).** *A Property Valuation Model for Rural Victoria*. (PhD Thesis) URL: [http://adt.lib.rmit.edu.au/adt/uploads/approved/adt\\_VIT20070221.150256/public/02whole.pdf](http://adt.lib.rmit.edu.au/adt/uploads/approved/adt_VIT20070221.150256/public/02whole.pdf).

**IAAO. International Association of Assessing Officers (2010).** *Standard on ratio studies*. page 60. URL: <http://www.iaao.org/uploads/Standard-on-Ratio-Studies.pdf>.

**Kauko, T. and d'Amato, M. (2009).** *Introduction: Suitability Issues in Mass Appraisal Methodology*. URL: <http://dx.doi.org/10.1002/9781444301021.ch1>.

**Kennedy, G.; Henning, S.; Vandever, L. and Dai, M. (1997).** Multivariate procedures for identifying rural land submarkets. *Journal of Agricultural and Applied Economics*, 29(2). URL: <http://econpapers.repec.org/RePEc:ags:joaacc:15066>.

**Orford, S. (2002).** Valuing locational externalities: a gis and multilevel modelling approach. *Environment and Planning B: Planning and Design*, 29(1):105–127. URL: <http://www.envplan.com/abstract.cgi?id=b2780>.

**Stumpf, M. (2007).** *Valor de inmuebles con fines fiscales*. Curso a Distancia: Aplicaciones del Catastro Multifinalitario en la definición de políticas de suelo urbano: 24.

**Williamson, I. P. (2001).** Land administration "best practice" providing the infrastructure for land policy implementation. *Land Use policy*, 18: 297-307.

**Witten, I. and Frank, E. (2005).** *Data Mining: Practical Machine Learning Tools and Techniques*. Second Edition. Morgan Kaufmann.

