

VARIABLES QUE INFLUYEN EN LA TASA DE PARO MUNICIPAL EN TIEMPOS DE CRISIS

Antonio Vaamonde Liste¹, Ricardo Luaces Pazos¹, Patricio Sánchez Bello¹ e Patricio Sánchez Fernández²

¹ Depto Estadística e Investigación Operativa, Universidade de Vigo

² Depto Economía Financiera y Contabilidad, Universidade de Vigo

RESUMO

Desde el comienzo de la crisis, a partir del año 2008, se ha observado una tendencia creciente en la tasa de paro, definida en el nivel municipal como el cociente entre el paro registrado y la población potencialmente activa. Esa tendencia se mantiene en el segundo semestre de 2013, aunque con un ritmo medio de crecimiento cada vez menor.

La tasa de paro en los municipios está asociada principalmente con la actividad económica, y su medida es importante para la definición de estrategias locales en la lucha contra el desempleo, basadas en el aprovechamiento de las ventajas específicas de la actividad local (turismo, artesanía, agricultura).

En este trabajo se analiza la relación entre la tasa de paro municipal y algunas magnitudes relacionadas con los niveles de actividad económica, con el objetivo de conocer mejor el problema del paro y las posibilidades de actuación para su corrección. Se utilizan técnicas estadísticas de regresión lineal múltiple, random forest, y redes neuronales SVM.

Palabras e frases chave: unemployment, regression, random forest, support vector machine.

1. INTRODUCCIÓN

Se utiliza la base de datos de La Caixa, disponible en serie temporal desde períodos muy anteriores a la crisis, con numerosas variables desagregadas a nivel municipal. El estudio se limita a los datos del último año completo disponible, 2011, ya que el análisis cronológico de la serie requiere probablemente una perspectiva más amplia que abarque la crisis en su conjunto, y deberá abordarse dentro de algunos años.

El número de municipios utilizados en la muestra es 145, todos los municipios españoles que superan los 50000 habitantes.

La variable dependiente es la tasa de paro municipal. Las variables explicativas son en algunos casos variables estructurales (población, proporción de inmigrantes, número de entidades financieras), y en otros casos la variación anual de algunas de estas variables (actividad industrial, automóviles, comercio minorista). Para la aplicación se emplea el programa estadístico R.

Variables explicativas:

Número de Tractores
Entidades de depósito
Población total
Superficies comercio alimentación
Variación de comercio minorista
Porcentaje de población extranjera
Variación de actividad industrial

Comercio de vestido y calzado
 Superficie de Centros Comerciales
 Variación de actividad de restauración
 Variación del número de camiones
 Variación del número de automóviles
 Porcentaje de mujeres

Se emplean diversas técnicas de regresión: lineal múltiple, Random Forest y Redes neuronales SVM. El objetivo es estudiar el valor predictivo de las variables explicativas, para conocer aquellas que tienen un mayor efecto sobre la tasa de paro municipal. La comparación de los tres modelos utilizados permitirá encontrar el más adecuado para los datos de paro.

Random Forest (Leo Breiman, 2001) consiste en una combinación de árboles de regresión en la que cada árbol se obtiene al azar de forma independiente mediante un vector aleatorio.

Support Vector Machines (Vapnik, 1995) es un modelo de clasificación supervisada, de carácter no probabilístico, que intenta separar las categorías lo máximo posible partiendo de una muestra de entrenamiento, y asigna los nuevos casos a la categoría más próxima.

Se comparan los tres métodos de predicción mediante el porcentaje de varianza explicada (coeficiente de determinación o pseudo R^2).

2. RESULTADOS

a) Regresión lineal múltiple

Los resultados de la aplicación del modelo de regresión lineal son los siguientes:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	7.389e+01	1.113e+01	6.636	7.75e-10	***
Tractores	-2.837e-04	7.634e-04	-0.372	0.71081	
Entidades.deposito	-6.313e-02	9.277e-03	-6.805	3.27e-10	***
Pob2011	3.332e-05	5.954e-06	5.595	1.23e-07	***
Sup.alimentacion	4.264e-05	8.386e-06	5.085	1.24e-06	***
Var.minoristas	-1.408e-01	3.882e-02	-3.627	0.00041	***
porc.extranj	-1.350e-01	2.889e-02	-4.673	7.26e-06	***
var.industria	-1.696e-02	2.284e-02	-0.742	0.45926	
vestido.calzado	5.606e-03	1.751e-03	3.202	0.00172	**
Sup.Centros.Comerciales	-6.274e-06	3.576e-06	-1.754	0.08172	.
Var.restauracion	-3.307e-02	3.920e-02	-0.843	0.40050	
Var.camiones	-9.990e-02	3.756e-02	-2.660	0.00880	**
Var.automoviles	1.151e-01	3.783e-02	3.043	0.00283	**
porc.mujeres	-1.231e+00	2.109e-01	-5.835	4.00e-08	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.395 on 131 degrees of freedom

Multiple R-squared: 0.575, Adjusted R-squared: 0.5328

F-statistic: 13.63 on 13 and 131 DF, p-value: < 2.2e-16

La mayoría de las variables explicativas utilizadas son significativas al nivel 0,01. El coeficiente de determinación supera el 50% de variabilidad explicada, lo que se puede considerar moderadamente satisfactorio.

Algunos coeficientes del modelo son positivos, y el valor alto de las variables correspondientes se asocia al incremento de la tasa de paro; así, mayor población significa mayor tasa de paro; los municipios más pequeños, en general menos desarrollados e industrializados, soportan mejor la crisis de desempleo.

Otros coeficientes son negativos, indicando una relación inversa: una mayor proporción de población femenina o extranjera está relacionada con tasas de paro más bajas.

b) Random Forest

Varianza explicada: 40,41409 %. Este método es menos predictivo que el modelo lineal.

Se representan en la figura 1 dos indicadores de importancia relativa de las variables explicativas. Una variable es más importante cuando su incorporación produce un cambio mayor en el error cuadrático medio, medida de precisión del ajuste. En el gráfico de la izquierda se muestra la importancia de cada una de las variables explicativas, en orden decreciente: la más importante es el porcentaje de población extranjera, con un efecto superior al 15% en el MSE en los casos utilizados para validación, y la menos importante población. El gráfico de la derecha muestra otro indicador de la importancia de las variables, basada en los cambios en el índice de Gini: un valor más elevado (porcentaje de extranjeros, entidades de depósito) indica que la variable explicativa produce nodos más consistentes y estables en los árboles de regresión.

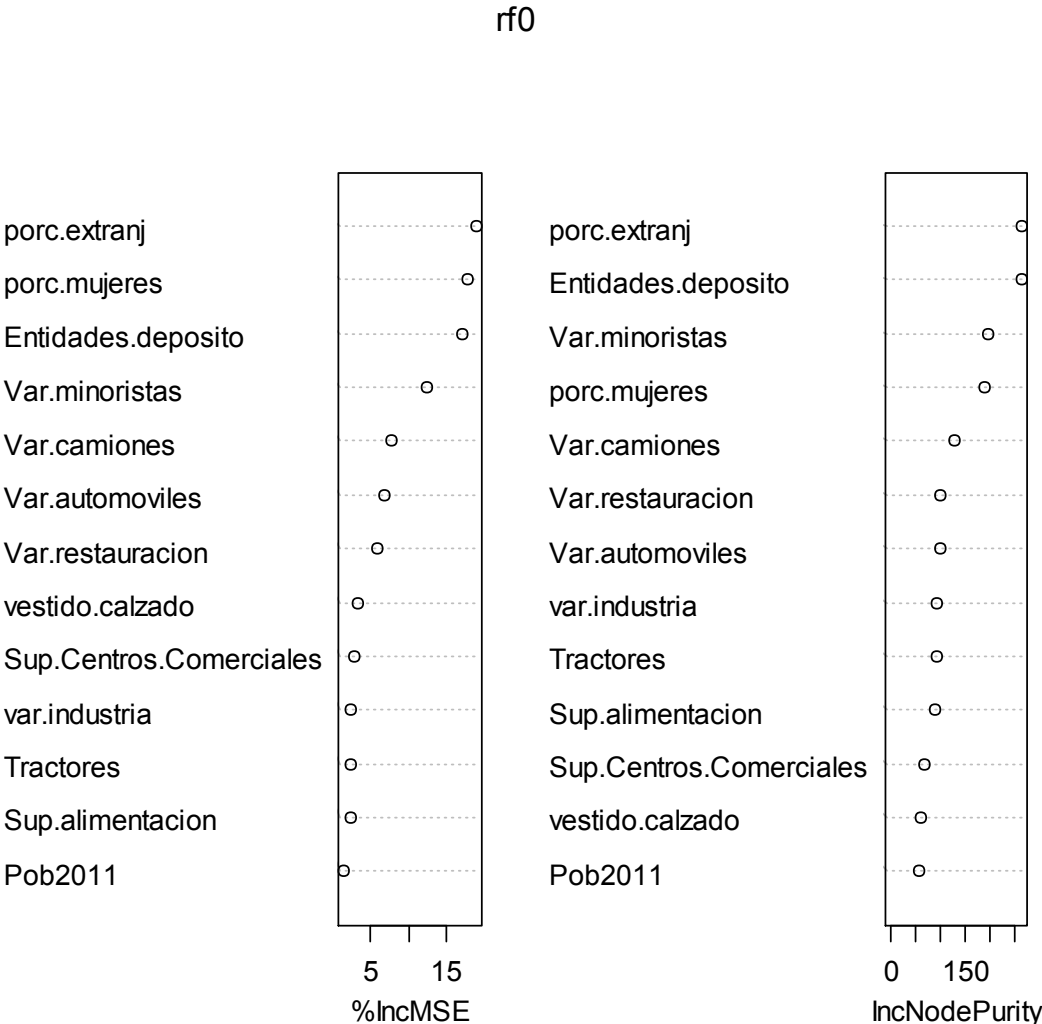


Figura 1: Random Forest. Importancia de las variables explicativas.

Se han utilizado 500 árboles para formar el “bosque”. Para valorar la convergencia del algoritmo, se comprueba en la figura 2 la estabilidad del error a partir de un número de árboles superior a 50.

tasa de error según número de árboles

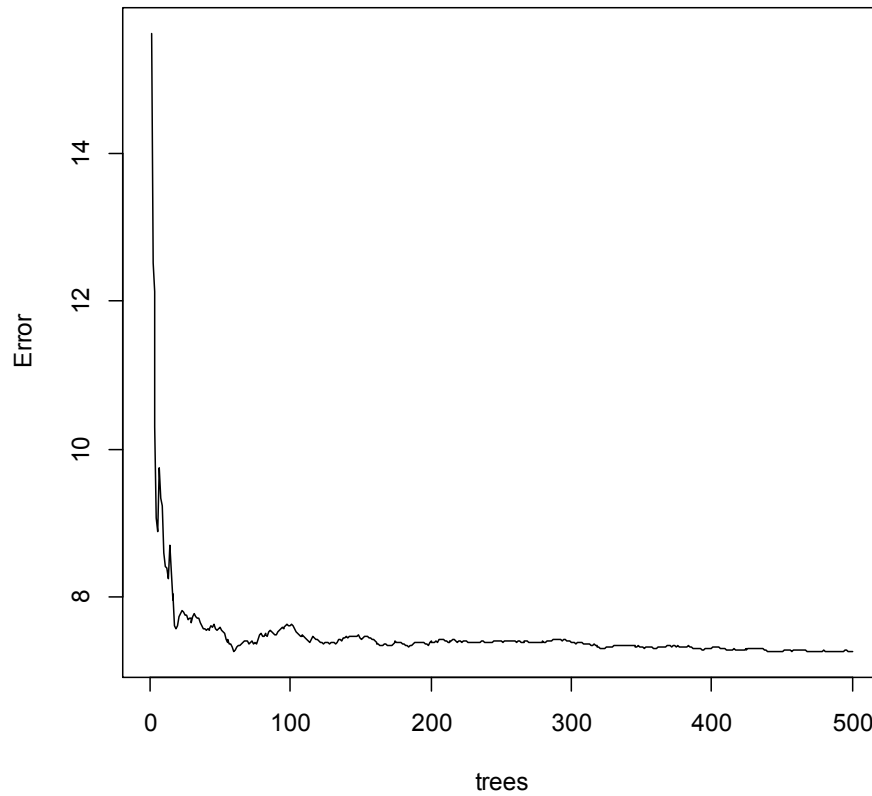


Figura 2: Random Forest. Tasa de error en función del número de árboles.

En el gráfico siguiente, figura 3, se muestra la tasa de paro comparada con la predicción en la totalidad de los municipios utilizados. El modelo parece ajustarse peor a los valores extremos de la tasa de paro (muy bajos o muy altos), lo que podría explicar la reducción observada en la varianza explicada: quizás esas tasas de paro anómalas responden a causas específicas o singulares, no modelizables, en los municipios afectados.

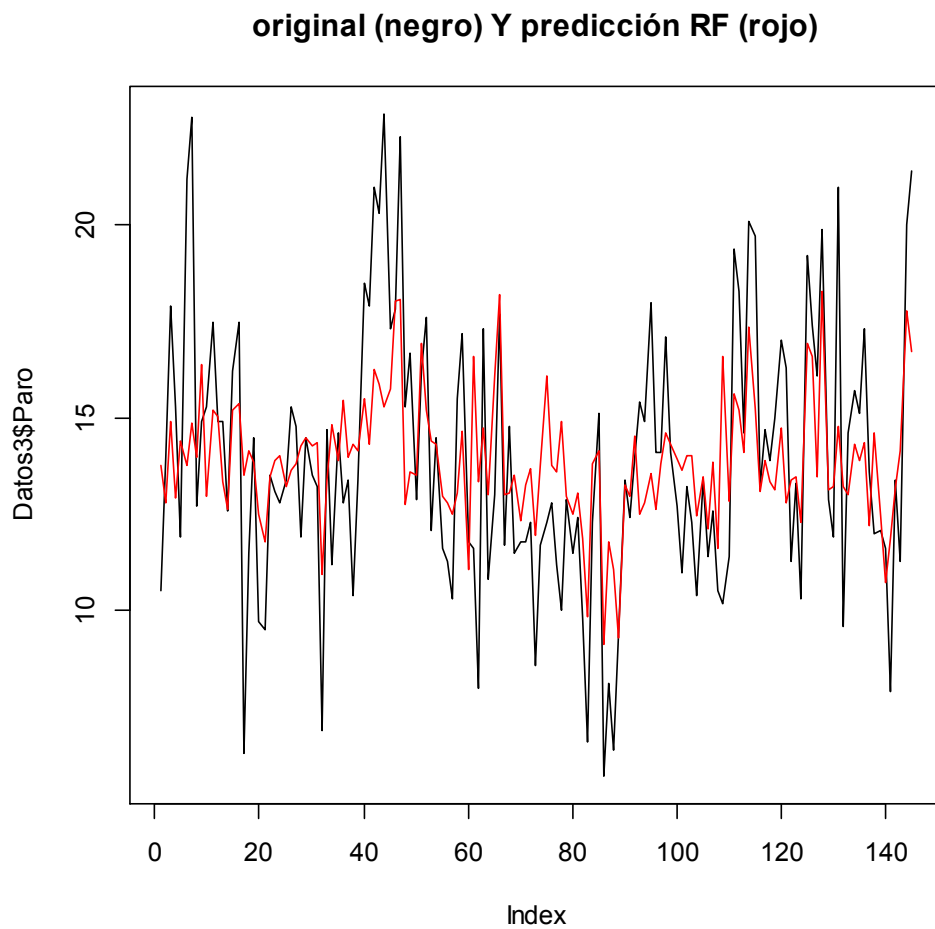


Figura 3: Random Forest. Predicción y valor observado en la tasa de paro municipal.

c) Support Vector Machine (SVM)

Varianza explicada: 69,56745 %. Es el método más predictivo de los tres utilizados. En el gráfico se muestra la predicción y el valor observado en la secuencia de municipios utilizados en la muestra.

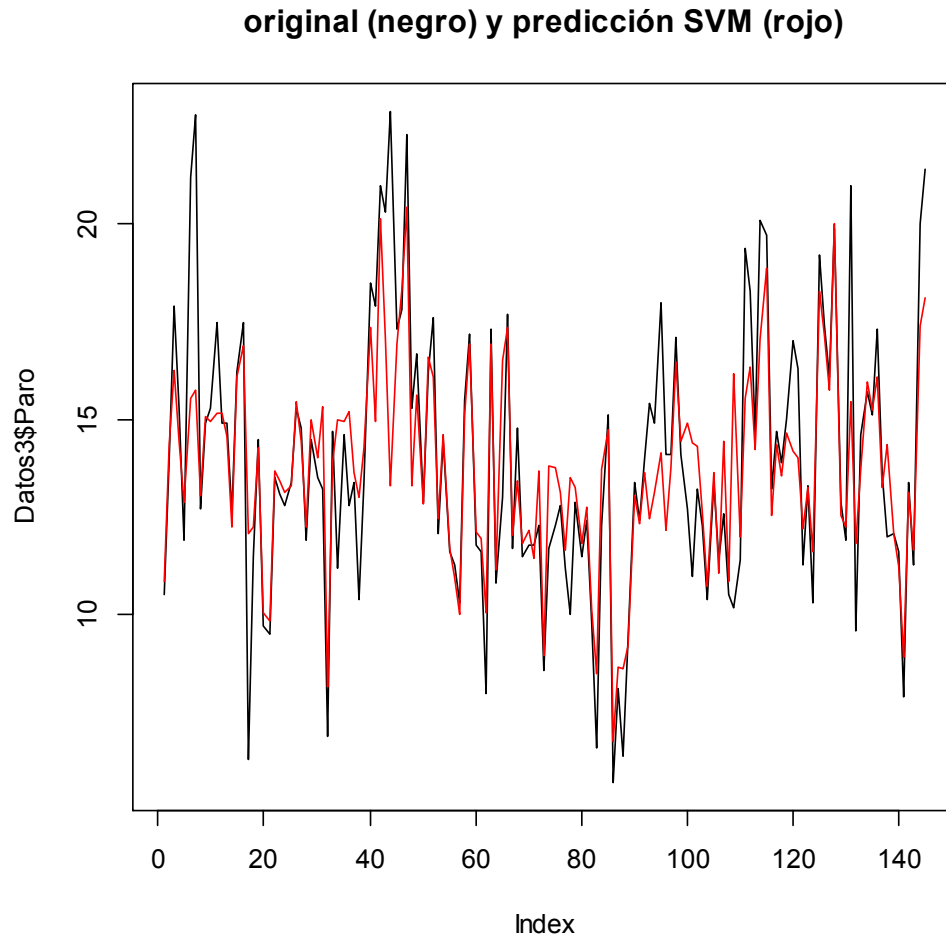


Figura 4: Support Vector Machine. Predicción y valor observado en la tasa de paro municipal.

El gráfico de dispersión que se muestra a continuación indica de nuevo la relación entre la variable observada y la predicción. Los valores más elevados de la tasa de paro muestran una mayor variabilidad y menor grado de ajuste.

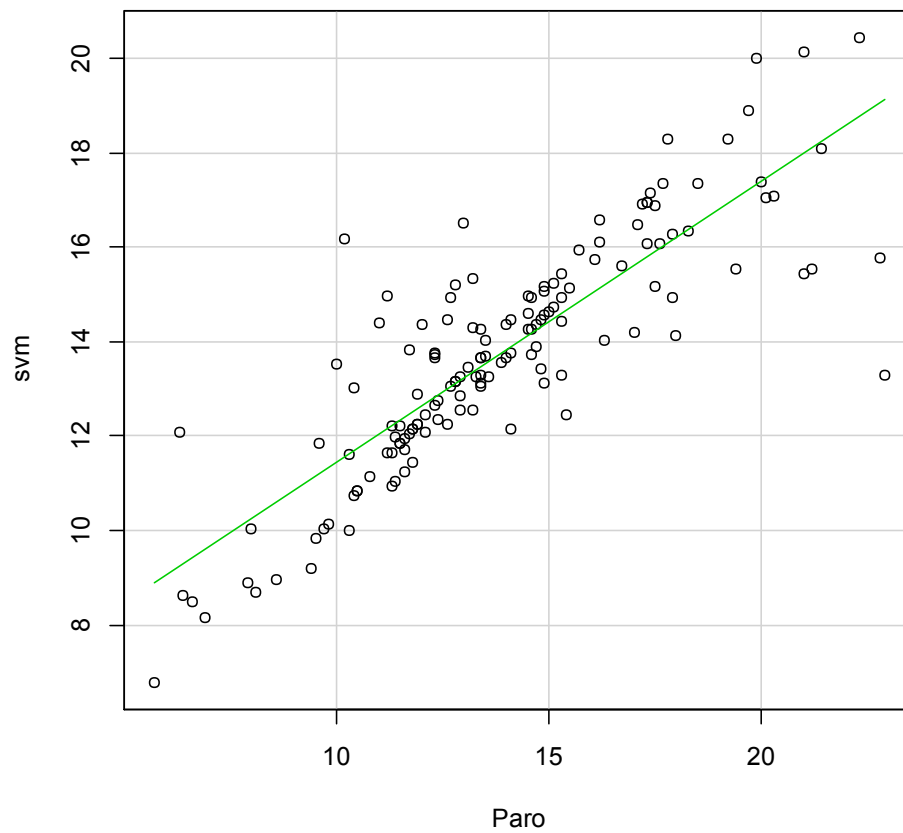


Figura 5: Support Vector Machine. Predicción (SVM) frente a valor observado.

REFERENCIAS

- Breiman, Leo (2001). Random Forests. *Machine Learning* 45 (1): pp. 5–32.
- Cortes, C. y Vapnik, V. (1995), Support-vector networks. *Machine Learning* 20 (3), 273–297.