UNIVERSIDAD NACIONAL DE ROSARIO

FACULTAD DE CIENCIAS EXACTAS, INGENIERÍA y AGRIMENSURA

CARRERA DE GRADO

LICENCIATURA EN CIENCIAS DE LA CUMPUTACION

Aplicación de modelos de regresión para la predicción de la rugosidad de pavimentos.

Autor: Regina Cecilia Muzzulini

Director: Dr. Rafael Namias

Codirector: Dr. Oscar Hugo Giovanon

2017

Dedicado a mis hijos Luca y Emma

Agradecimientos

Este trabajo de tesina realizado en el laboratorio vial del IMAE-FCEIA-UNR es un esfuerzo en el cual, directa o indirectamente, participaron varias personas opinando, corrigiendo, teniéndome paciencia, dandome ánimos.

Quiero agradecer a mi director, Dr. Rafael Namias por su constante seguimiento y dedicación. A mi codirector Dr. Oscar Giovanon por la motivación y el apoyo recibido a lo largo de estos años.

A mis compañeras y amigas de estudio Erica, Fernanda y Maura, por haberme acompañado a lo largo de la carrera, compartiendo apuntes, tiempo, mates, momentos para distenderse después de horas de estudio.

Y, por supuesto, el agradecimiento más profundo y sentido es para mi familia. Sin su apoyo, colaboración e inspiración habría sido imposible llevar a cabo esta dura tarea.

En especial, a mis padres, por haberme dado la oportunidad de estudiar en su momento y por su ejemplo de lucha y honestidad.

A mi esposo Ariel, por creer en mí, por su paciencia, su generosidad, apoyo incondicional y por compartir todos mis proyectos.

A mis hijos Luca y Emma, por ser tan dulce y darme la fuerza necesaria para seguir adelante...

A todos muchas gracias!!!

Resumen

El relevamiento periódico de la condición del pavimento, en forma ordenada y sistemática, así como también la evolución de los deterioros del mismo, permiten conformar la función del comportamiento de los distintos tramos de una ruta; para poder predecir las tareas de mantenimiento necesarias en magnitud y oportunidad.

Si bien resulta conveniente realizar este relevamiento desde el inicio de la vida de la estructura, determinando así el nivel cero o punto de partida de la condición; no siempre se dispone de la totalidad de la información.

Existen diferentes parámetros (deterioros) para la valoración del comportamiento de la superficie de la carretera: rugosidad, ahuellamiento, fisuras, baches, desprendimientos, exudación, etc. La característica que se pretende evaluar en esta tesina para determinar el estado de deterioro del pavimento es la rugosidad, que afecta en forma muy importante la dinámica del vehículo, la calidad de circulación, el efecto dinámico de las cargas y el drenaje. Los defectos de rugosidad son percibidos por los usuarios como movimientos vibratorios que afectan el confort de circulación y constituye el principal parámetro para la estimación del costo de los usuarios, cuya disminución determina la factibilidad de la realización de obras de mejora sobre el pavimento.

Cuando los indicadores de estado superficial alcanzan determinados niveles de alerta, se deben proyectar las correspondientes tareas de mejoras; posibilitando que las obras puedan ser realizadas en el momento adecuado sin que se incrementen los costos de mantenimiento.

En esta tesina se presentan técnicas de “aprendizaje automatizado” para predecir la

evolución de los valores de rugosidad. Se intentó modelizar el problema de *degradación del pavimento* a partir de modelos estadísticos que permiten ajustar un modelo de regresión para predecir la evolución de los valores de rugosidad a partir de técnicas del estado del arte.

Índice general

Índice de figuras

Índice de tablas

Introducción

¿Qué es un pavimento?

Desde el punto de vista ingenieril, el pavimento consiste en una estructura formada por varias capas sobre la que actúan cargas en superficie, que deben ser capaces de trasmitir durante su vida útil las tensiones en profundidad, de tal forma que no superen las tensiones y deformaciones específicas admisibles, tanto en el suelo de fundación como en cada una de sus capas.

Para el usuario común, el pavimento es una superficie que debe permitir la circulación del tránsito mixto, en condiciones de seguridad y comodidad, bajo cualquier condición climática, durante un tiempo prolongado.

La problemática que se plantea, una vez que el pavimento ha sido puesto en servicio, es evaluar cuál es la respuesta del mismo frente a las acciones de tránsito y clima, y realizarlo con cierta periodicidad, determinando si son favorables o desfavorables. A esto se lo conoce como *auscultación* del estado de la carretera; y consiste en la valoración del estado de los deterioros superficiales y estructurales que afectan la calidad del servicio brindado a los usuarios.

¿Porqué es tan importante la auscultación del estado del pavimento?

Porque a partir de la misma es posible conformar la función del comportamiento de los distintos tramos de una ruta; para poder predecir las tareas de mantenimiento necesarias en magnitud y oportunidad, y también obras de mejoras.

Existen diferentes parámetros de control de la calidad, de seguridad y del confort de la comodidad, como lo es la *rugosidad*. Los defectos de la rugosidad son percibidos por los

usuarios como movimientos vibratorios que afectan el confort de circulación y constituye el principal parámetro para la estimación del costo de los usuarios.

Al momento de priorizar obras o mantenimientos de rutina es necesario el desarrollo de herramientas adecuadas que permitan el análisis de la evolución de deterioros, posibilitando que puedan ser realizadas en el momento adecuado sin que se incrementen los costos de mantenimiento [1].

En este trabajo se presentan técnicas de "aprendizaje automatizado" para predecir la evolución de los valores de rugosidad.

El aprendizaje automatizado (*Machine Learning*) es el subcampo de las *ciencias de la computación* y una rama de la *inteligencia artificial* que tiene por objetivo desarrollar técnicas que permitan a las computadoras *aprender*.

"Un sistema se dice que es capaz de aprender de la **experiencia (E)** con una serie de **tareas (T)** y una medida del **rendimiento (P)** si su desempeño en las tareas **T** mejora con **E**" [2].

En el modelado de datos empíricos se utiliza un proceso de inducción para construir un modelo del sistema, del que se espera deducir las respuestas del sistema que aún no se han observado. La cantidad y calidad de las observaciones rigen el desempeño de este modelo empírico.

El análisis actual del problema se focalizó en el *aprendizaje supervisado*, cuyo objetivo es crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos. Realiza predicciones de cosas desconocidas basadas en comportamientos o características que se han visto en los datos ya almacenados.

Como la rugosidad es una variable de tipo continua; donde el rango de la misma define una escala que va desde 0 para superficies idealmente planas (las pistas de aeropuertos pueden brindar valores inferiores a la unidad), tomando valores entre 1 y 2 m/km para pavimentos recién construídos, y valores superiores a 6 m/km para superficies notoria-

mente deterioradas, se intenta modelizar el problema a partir de modelos estadísticos que permiten ajustar un modelo de regresión para predecir la evolución de la misma, a partir de técnicas del estado del arte.

¿Qué algoritmo de aprendizaje automático usar?
El algoritmo a emplear depende del tamaño, la calidad y la naturaleza de los datos. La idea fundamental del *aprendizaje automatizado* es encontrar patrones que puedan generalizarse, para poder aplicar esta generalización sobre los casos que todavía no se han observado y realizar predicciones. Puede ocurrir que durante el entrenamiento se descubran *casualidades* en los datos que se parecen a patrones interesantes, pero que no generalicen. Esto se conoce con el nombre de *sobreajuste*.

Todos los modelos del *aprendizaje automatizado* tienen tendencia al *sobreajuste*, por lo que hay que tratar de tomar medidas preventivas para reducirlo lo más posible. Una de las principales estrategias es la *validación cruzada*.

Partiendo de un conjunto de ejemplos $\mathcal{C}$ como $(x_1, y_1), ..., (x_n, y_n)$, de los cuales se conocen una serie de variables, una vez determinada la variable respuesta y las variables explicativas, se buscan procedimientos para construir un *modelo* para predecir el valor de y para un nuevo valor de x, comprendiendo la relación entre x e y. El objetivo de la regresión es minimizar el error entre la función aproximada y el valor de la aproximación. Si bien en este estudio, no es necesario obtener una respuesta lo más precisa posible, con una aproximación ya es útil.

La cantidad de características puede ser muy grande en comparación con la cantidad de puntos de datos. Una gran cantidad de características puede trabar algunos algoritmos de aprendizaje y provocar que el tiempo de entrenamiento sea demasiado largo. Por lo que una mala modelización puede llevar a superposición de datos.

Para el desarrollo de las técnicas de prueba se utiliza como lenguaje a **Python**. Y como paquetes de software que incluyen algoritmos de aprendizaje automatizado, **Scikit-learn**. Esta librería también facilita las tareas de evaluación, diagnóstico y validaciones cruzadas ya que proporciona varios métodos de fábrica para poder realizar estas tareas en forma

muy simple [3][4].

Se espera que esta herramienta pueda ser implementada en sistemas de gerenciamiento viales.

Para poder cumplimentar con estos objetivos, la presente tesina está organizada en cinco capítulos.
En el Capítulo 2, se introduce al lector en forma teórica, desarrollando conocimientos en los modelos de regresión como lo son *Support Vector Machine* y *Random Forest*.

El Capítulo 3 se destina a la presentación de los enfoques para abordar el problema de la auscultación de pavimentos para la valoración del comportamiento de la superficie de la carretera: rugosidad, ahuellamiento, fisuras, baches, desprendimientos, exudación, etc.

El Capítulo 4 se basa en la aplicación de las metodologías propuestas de modelos de regresión. Partiendo de un análisis básico, mejorando los conjuntos para optimizar la estrategia.

Por último, el Capítulo 5, se dedica a la presentación de las consideraciones finales del estudio y las propuestas para la continuación del mismo, con diferentes líneas de investigación.

Modelos de regresión

2.1. Support Vector Machine

La idea de *Support Vector Machine* es seleccionar un hiperplano de separación que equidiste de los ejemplos más cercanos de cada clase para conseguir lo que se denomina un margen máximo a cada lado del hiperplano. Para definir el hiperplano, se consideran sólo los ejemplos de entrenamiento que distan del hiperplano la distancia margen. Estos ejemplos reciben el nombre de vectores soporte.

Se comenzará con una breve descripción de los casos de máquinas de vectores soporte para la clasificación para luego poder abordar mejor los vectores soporte para regresión.

Tanto para el caso de *Support Vector Machine* como *Support Vector Machine Regression*, el objetivo es realizar la predicción a partir de un problema de optimización geométrica que se puede escribir como un problema de optimización cuadrático convexo con restricciones lineales, en principio resoluble mediante cualquier procedimiento de optimización no lineal. Se tratará de ver cómo se puede obtener dichos problemas y las transformaciones necesarias para facilitar su resolución.

2.1.1. Caso linealmente separable

Para un conjunto linealmente separables de ejemplos $(x_1, y_1), ..., (x_n, y_n)$, donde $x_i \in \mathbb{R}^d$ e $y_i \in \{+1, -1\}$ se define el hiperplano de separación como una función lineal capaz de

separar dicho conjunto:

$$D(x) = (w_1x_1 + w_2x_2 + ... + w_dx_d) + b = \langle w, x\rangle + b,$$

donde $w_i \in \mathbb{R} \quad \forall i = 1, ..., d$ y $b \in \mathbb{R}$.

Este hiperplano deberá cumplir con la siguiente desigualdad:

$$\langle w, x\rangle + b \geq 0 \quad si \quad y_i = +1 \quad \forall i = 1, ..., n$$

$$\langle w, x\rangle + b \leq 0 \quad si \quad y_i = -1 \quad \forall i = 1, ..., n$$

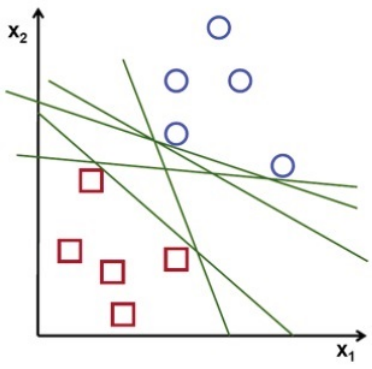
que sería equivalente a:

$$y_i(\langle w, x\rangle + b) \geq 0 \quad \forall i = 1, ...n$$

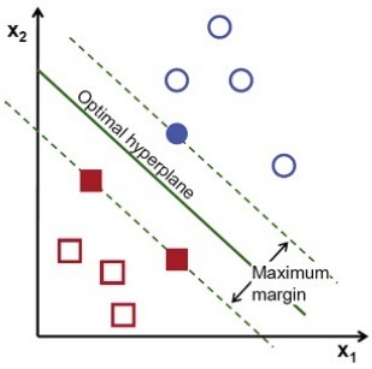
Pero el hiperplano:

$$y_iD(x_i) \geq 0 \quad \forall i = 1, ...n$$

que intenta separar las dos clases no suele ser único. Se introduce el concepto de *margen* τ, que consiste en la distancia mínima entre dicho hiperplano y el ejemplo más cercano a cada clase.



(a) Se aprecia la NO unicidad de la solución.



(b) Una vez impuesto el margen máximo, el hiperplano es único.

Sea el ejemplo x'. La distancia de dicho ejemplo al hiperplano viene dada por:

$$\frac{|D(x')|}{\|w\|} \tag{2.1.1}$$

Tomando en cuenta el *margen* y la desigualdad (2.1.1), se deduce que encontrar el hiperplano óptimo, es equivalente a encontrar el valor de w que maximiza el *margen*:

$$y_i D(x_i) \geq \tau \|w\| \quad \forall i = 1, ...n$$

Maximizar el *margen* es equivalente a minimizar $\|w\|$

Se puede reescribir de forma equivalente como un problema de programación cuadrática de la siguiente forma:

$$\text{Minimizar} \quad \frac{1}{2}\|w\|^2$$

$$\text{sujeto a:} \quad y_i(\langle w, x_i \rangle + b) - 1 \geq 0 \quad i = 1, ..., n$$

Cuyo problema dual asociado es:

$$\text{Maximizar} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle (x_i), (x_j) \rangle$$

$$\text{sujeto a:} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\alpha_i \leq 0 \quad i = 1, ..., n$$

El hiperplano obtenido luego de resolver este problema se denomina hiperplano de separación duro o *hard margin*.

2.1.2. Caso lineal no separable

Hallar un conjunto con dos clases totalmente separables es poco probable, entre otras cosas por la existencia de *ruido* en los datos. Se introduce así, un conjunto de variables

reales y positivas denominadas *variables artificiales*, ξ_i, $i = 1, ..., n$, permitiendo ejemplos no separables:

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \qquad \xi_i \geq 0, i = 1, ..., n$$

La función a optimizar debe incluir las *variables artificiales* para controlar el error en la clasificación. Así, el nuevo problema de optimización a resolver quedaría:

$$\text{Minimizar} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\begin{aligned} \text{sujeto a:} \quad & y_i(\langle w, x_i \rangle + b) + \xi_i - 1 \geq 0 \qquad i = 1, ..., n \\ & \xi_i \geq 0 \qquad i = 1, ..., n \end{aligned}$$

donde C es una constante positiva a determinar de la que puede depender el clasificador.

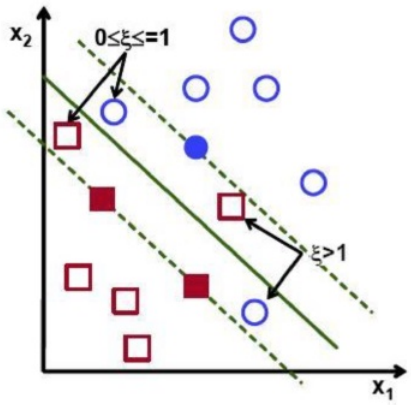


Figura 2.1: Ejemplo de problema no linealmente separable.

Cuyo problema dual asociado es:

$$\text{Maximizar} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \langle (x_i), (x_j) \rangle$$

$$\begin{aligned} \text{sujeto a:} \quad & \sum_{i=1}^{n}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \qquad i = 1, ..., n \end{aligned}$$

El hiperplano definido tras resolverlo se denomina hiperplano de margen blando o *soft*

margin.

2.1.3. Caso no lineal

Es frecuente, en los problemas reales, clasificar un conjunto donde los ejemplos no suelen ser separables. Para usar estrategias no lineales, se realiza la inmersión no lineal del conjunto, en un espacio de dimensión mayor donde sí sean separables. Y en este nuevo espacio, *espacio de características*, se busca el hiperplano de separación.

Sea $\phi : \mathbb{X} \to \mathcal{F}$ la función que hace corresponder a cada punto de entrada x un punto en el espacio de características $\mathcal{F}$. Por lo tanto, se debe hallar el hiperplano de separación en este nuevo espacio $\mathcal{F}$. Este hiperplano en el espacio de características se transforma en una función no lineal que separa el conjunto en el espacio original de entradas. La función de decisión en el espacio de caracterísiticas viene dada por:

$$D(X) = \langle w, \phi(x) \rangle + b$$

El problema de optimización asociado es similar al descrito para el caso *lineal no separable*, con la diferencia que el hiperplano de separación se determina en el espacio de características, es decir:

$$\text{Minimizar} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{2.1.2}$$

$$\begin{aligned} \text{sujeto a:} \quad y_i(\langle w, \phi(i) \rangle + b) + \xi_i - 1 \geq 0 \qquad & i = 1, ..., n \\ \xi_i \geq 0 \qquad & i = 1, ..., n \end{aligned}$$

La complejidad de (2.1.2) depende de la dimensión del espacio de características. Por esto, se considera el problema dual asociado, cuya complejidad depende del número de ejemplos:

$$\text{Maximizar} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle$$

$$\text{sujeto a:} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C \qquad i = 1, ..., n$$

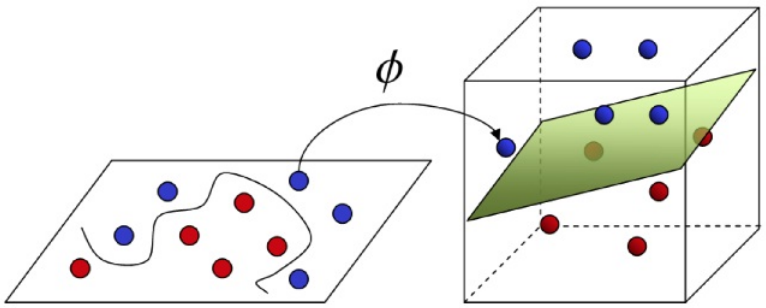


Figura 2.2: Ejemplo gráfico de una posible función ϕ.

En el problema anterior no es necesario conocer las componente ϕ_i, sino los productos escalares $\langle \phi(x_i), \phi(x_j) \rangle$. Para resolver este tipo de problemas se usa el *Truco del Kernel*, donde:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

De esta manera se puede construir una función de decisión en el espacio original sin usar explícitamente la función de inversión ϕ [5].

2.2. Support Vector Regression

La técnica de los vectores soportes es una herramienta universal para resolver problemas de estimación de funciones multidimensionales. Dado el buen resultado en el ámbito de la clasificación se plantea un problema similar para abordar la regresión.

Se trata de seleccionar el hiperplano regresor que mejor se ajuste al conjunto de datos de entrenamiento. La idea se basa en considerar una distancia margen ε, de manera que todos los ejemplos se encuentren en una *banda* o *tubo* entorno a dicho hiperplano. Es decir, que disten a una cantidad menor de ε del hiperplano. A la hora de definir el hiperplano, solo se consideran los ejemplos que disten más de ε del hiperplano. Estos ejemplos serán

los considerados *vectores soporte*.

En *Support Vector Machine* para clasificación, se usa el márgen para determinar la separación entre las dos clases de puntos. A mayor margen, mayor seguridad de que se está ante un hiperplano de separación bueno.

En la regresión, se pretende que la función esté lo más próxima posible a los puntos. La formación de una *banda* o *tubo* alrededor de la verdadera función de regresión sería el equivalente de lo que sería para la clasificación el *margen*. Los puntos no contenidos dentro del tubo se identifican con la posibilidad estricta de las variables artificiales asociadas, que cuantifican el error cometido entre la aproximación y el valor real de cada ejemplo del conjunto de entrenamiento.

2.2.1. Caso lineal

El método *Support Vector Machine* para regresión es una aplicación de las *Support Vector Machine* en la aproximación de funciones. Tambien denominada ε-SV, su objetivo es: dado un conjunto finito de datos $\{(x_i, y_i)\}$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, $i = 1, ..., d$, encontrar una función de regresión, en principio lineal, de la siguiente forma:

$$f(x) = (w_1 x_1 + ... + w_d x_d) + b = \langle w, x \rangle + b \tag{2.2.1}$$

donde $w_i \in \mathbb{R} \quad \forall i = 1, ..., d$ y $b \in \mathbb{R}$.
La función f debe estar como máximo a ε desviaciones de los objetivos y_i para todos los datos de entrenamiento x_i y a la vez sea tan "plana" como sea posible. En este contexto, "plana" significa que se quiere encontrar una w *pequeña*. Para esto, se minimiza la norma euclídea $\|w\|^2$.

Se transforma el problema (2.2.1) en uno de optimización convexa:

$$\text{Minimizar} \quad \frac{1}{2}\|w\|^2 \tag{2.2.2}$$

$$\text{sujeto a:} \quad y_i - \langle w, x_i \rangle - b \leq \varepsilon$$
$$-y_i + \langle w, x_i \rangle + b \leq \varepsilon$$

La suposición tácita del problema anterior es tal que la función f actualmente existente, aproxima todos los pares (x_i, y_i) con precisión ε; en otras palabras, que el problema sea *factible*. Sin embargo, es posible que se desee permitir algunos errores. Se introduce una función de pérdida *soft margin*, introduciendo variables de holguras. Estas variables positivas y reales ξ_i, ξ_i^* cuantifican el error cometido entre la aproximación y el valor real de cada ejemplo del conjunto de entrenamiento, ignorando los errores que están situados dentro de la distancia determinada del valor real.

Se reformula el problema de optimización convexa (2.2.2) quedando:

$$\text{Minimizar} \quad \frac{1}{2}\|w\|^2 + C\sum_i^n (\xi_i + \xi_i^*)$$

$$\begin{aligned} \text{sujeto a:} \quad y_i - \langle w, x_i \rangle - b &\leq \varepsilon + \xi_i \qquad i = 1, ..., n \\ -y_i + \langle w, x_i \rangle + b &\leq \varepsilon + \xi_i^* \qquad i = 1, ..., n \\ \xi_i, \xi_i^* &\geq 0 \qquad i = 1, ..., n \end{aligned}$$

Donde la constante $C \geq 0$, llamada *constante de regularización*, determina el compromiso entre lo "plano" de f y la cantidad de desviaciones mayores a ε que están permitidas. La fórmula descrita anteriormente corresponde a la relación la cual se denomina *función de pérdida - épsilon intensivo* descrita por:

$$L = \begin{cases} 0 & \text{si} \quad |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon & \text{en caso contrario} \end{cases}$$

La siguiente figura representa la situación gráficamente. Los puntos fuera de la región sombreada contribuye a un costo, es decir, las variables miden el costo de los errores en los puntos de entrenamiento. De manera que si el valor predicho está dentro del tubo, la pérdida es cero, y si el valor predicho está fuera, la pérdida es la diferencia entre el valor

predicho y el radio ε del tubo.

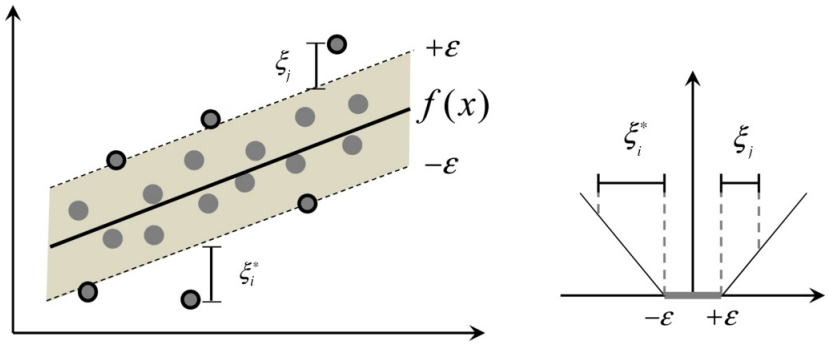


Figura 2.3: Ejemplo de cómo englobar a todos los ejemplos en una *banda* entorno a la función de predicción f.

Con el uso de esta función de pérdida se asegura existencia global y al mismo tiempo la optimización de la generalización.

Tras la obtención del problema primal, se pasa a plantear el problema dual asociado. La idea es utilizar un método estándar de dualización. Utilizando la función de *Lagrange* con la función objetivo y las restricciones correspondientes mediante la introducción de un conjunto de variables duales:

$$\text{Maximizar} \quad -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)\langle x_i,x_j\rangle-\varepsilon\sum_{i=1}^{n}(\alpha_i^*+\alpha_i)+\sum_{i=1}^{n}y_i(\alpha_i^*-\alpha_i) \quad (2.2.3)$$

$$\text{sujeto a:} \quad \sum_{i=1}^{n}(\alpha_i^*-\alpha_i)=0$$

$$\alpha_i^*,\alpha_i\in[0,C]$$

La solución del problema (2.2.3) está dada por la función de predicción:

$$f(x)=\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)\langle x_i,x\rangle+b$$

De esta manera se obtiene la función buscada sin depender la resolución del problema de la dimensión en la que se encuentran los ejemplos de entrada, dependiendo únicamente

de los *vectores soporte* [6][7][8].

2.2.2. Caso no lineal

La formulación (2.2.3) se puede generalizar al caso no lineal, que se requiere para modelar adecuadamente los datos, sustituyendo el producto escalar por una función núcleo k.

Sea el conjunto $\{(x_i, y_i), ..., (x_n, y_n)\}$ donde $x_i \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$. Sea $\phi : \mathcal{X} \to \mathcal{F}$ la función que hace corresponder a cada punto de entrada x un punto en el espacio de características $\mathcal{F}$. Es decir, la entrada x está primero mapeado a un $d - dimensional$ espacio de características utilizando algun mapeo (no lineal) fijo, y luego un modelo lineal se construye en este espacio de características. Este espacio de características puede ser de dimensión elevada o incluso infinita.

La idea es trasladar los ejemplos a este nuevo espacio de características y hallar la función que mejor aproxime las imágenes del conjunto. Se plantea el problema primal al igual que en el caso lineal a diferencia que en este caso no depende de los ejemplos del conjunto, sino de sus imágenes por una cierta función ϕ:

$$\text{Minimizar} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2.2.4}$$

$$\begin{aligned} \text{sujeto a:} \quad y_i - \langle w, \phi(x_i)\rangle &\leq \varepsilon + \xi_i \qquad i = 1, ..., n \\ -y_i + \langle w, \phi(x_i)\rangle &\leq \varepsilon + \xi_i^* \qquad i = 1, ..., n \\ \xi_i, \xi_i^* &\geq 0 \qquad\qquad i = 1, ..., n \end{aligned}$$

Donde ξ_i o ξ_i^* es cero si el punto de la muestra está dentro del tubo. Si el punto observado está por encima del punto, ξ_i es la diferencia positiva entre el valor observado y ε. Similarmente, ξ_i^* será diferente de cero si el punto observado está por debajo del tubo. Al menos que el punto se encuentre dentro del tubo, en cuyo caso, ambas restricciones serán cero.

La nueva *función de pérdida - épsilon intensivo* va a estar definida por:

$$L = \begin{cases} 0 & \text{si} \quad |y - f(X)| \leq \varepsilon \\ |y - f(X)| - \varepsilon & \text{en caso contrario} \end{cases}$$

En este caso, la función que se está buscando será de la siguiente forma:

$$f(X) = \langle w, \phi(x) \rangle + b$$

Donde ϕ denota entonces un conjunto de transformaciones no lineales y b es el sesgo.

Como ya se ha mencionado anteriormente, la complejidad del problema (2.2.4) depende de la dimensión en la que se encuentran los ejemplos, tras ser transformados por la función ϕ. Estos ejemplos podrían tener una dimensión muy alta, lo que complicaría la resolución del problema primal. Por esto, se considera el problema dual asociado:

$$\text{Maximizar} \quad -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)\langle\phi(x_i),\phi(x_j)\rangle-\varepsilon\sum_{i=1}^{n}(\alpha_i^*+\alpha_i)+\sum_{i=1}^{n}y_i(\alpha_i^*-\alpha_i) \quad (2.2.5)$$

$$\text{sujeto a:} \quad \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) = 0$$

$$\alpha_i, \alpha_i^* \in [0, C]$$

La función objetivo solo depende del producto escalar de las imágenes de los ejemplos. El algoritmo del *Truco del Kernel* es ampliamente utilizado en los algoritmos de cálculo de productos escalares de la forma $\langle \phi(x), \phi(x') \rangle$ en espacio características $\mathcal{F}$

2.2.3. Análisis de las variables

La fórmula (2.2.5) permite deducir varias conclusiones:

- Los multiplicadores Lagrangianos α_i, α_i^* no pueden ser simultáneamente diferentes de cero, por lo tanto, la restriccíon $\alpha_i \alpha_i^* = 0$ se cumple.

Es decir, no se pueden activar a la vez las dos variables duales asociadas a un mismo ejemplo.

- Aquellos puntos x_i en los cuales el error de interpolación es mayor o igual a ε son los denominamos *vectores soporte*. Los puntos en los que el error de interpolación es menor que ε nunca son *vectores de soporte*, y no forman parte de la solución. Podrían eliminarse del conjunto de datos, y si se volviera a resolver el problema de programación sobre el conjunto reducido, se encuentraría la misma solución.
- Solo los ejemplos (x_i, y_i) tales que $\alpha_i = 0$ ó $\alpha_i^* = 0$ quedarían fuera del *tubo* o *banda* construido.
- En los casos que $\alpha_i, \alpha_i^* \in (0, C)$, se tiene que las variables ξ_i, ξ_i^* correspondientes, se deben anular.

Si C es grande, se minimiza el error cuadrático medio en el conjunto de entrenamiento que puede dar pobres generalización a un conjunto de *test*. Se puede hallar un buen valor de C variándolo para encontrar el mejor rendimiento en un conjunto de validación y puede aplicar dicho C para el conjunto de *test*. C es muy útil si la dimensionalidad del conjunto de características es mayor que el número de ejemplos.

El parámetro ε controla el ancho de la zona ε-*insensible*, usada para ajustar los datos de entrenamiento. El valor de ε puede afectar el número de *vectores soportes* utilizados para construir la función de regresión. Cuánto más grande ε, se seleccionan pocos *vectores soportes*. Por otra parte, un valor grande ε da como resultado estimaciones más "planas". Por lo tanto, C y ε afectan la complejidad del modelo, pero de manera diferente.

La precisión de la estimación depende de una buena selección de los parámetros C, ε y los parámetros del *kernels*. El problema de la selección de parámetros óptimos se complica aún más por el hecho de que la complejidad del modelo (y por lo tanto su rendimiento) dependen de estos tres parámetros.

2.2.4. Ventajas

Una de las ventajas de la utilización de *Vector Support Machine* para la regresión es que pueden ser utilizadas para evitar dificultades usando funciones lineales en el espacio de características de mayor dimensión; y la optimización del problema es transformado en problemas duales cuadráticos convexos. Las funciones de pérdida que se utilizan para penalizar los errores que son mayores al umbral ε, conducen a la escasa representación de la regla de decisión, dando ventajas algorítmicas significativas y de representación [5][6][7][8][9].

2.3. Random Forest

Random Forest, introducidos por Leo Brieman (2001) son una extensión de la idea del *bagging* de Breiman [10]. Se pueden utilizar para una variable de respuesta categórica como clasificación o continua. Del mismo modo, las variables predictoras pueden ser categórica o continuas.

Random Forest son una combinación de árboles predictores de tal manera que cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para todos los árboles en el bosque. El error de generalización para los bosques converge a un límite a medida que el número de árboles en el bosque se hace grande.

En un árbol de clasificación estándar, la idea es dividir el conjunto de datos, basado en la homogeneidad de los datos. Un árbol de decisión se construye de arriba (nodo raíz) hacia abajo, e implica particionar los datos en subconjuntos que contienen instancias con valores similares (homogéneos).

Random Forest Regression son un conjunto de árboles de regresión diferentes y se utilizan para la regresión múltiple no lineal, donde cada hoja contiene una distribución para la variable de salida continua s.

Se intentará abordar primero los bosques aleatorios para la clasificación, para luego comprender mejor los bosques aleatorios para la regresión. Veremos algunos antecedentes teóricos para abordar mejor los bosques aleatorios. Se intentará demostrar que la exactitud de un bosque al azar depende de la fuerza de los clasificadores de árboles individuales y una medida de la dependencia entre ellos. Se introducirá los bosques usando la selección aleatoria de características en cada nodo para determinar la división.

2.3.1. Definición

Para un *k-ésimo árbol*, se genera un vector aleatorio Θ_k, independientemente de los anteriores vectores aleatorios $\Theta_1, ..., \Theta_{k-1}$, pero con la misma distribución. Se hace crecer un árbol usando el conjunto de entrenamiento y Θ_k, resultando en un clasificador $h(x, \Theta_k)$ donde x es un vector de entrada. Después de que se genera un gran número de árboles, votan por la clase más popular.

Formalmente, un bosque aleatorio es un clasificador que consiste en una colección de clasificadores estructurados en árbol $\{h(x, \Theta_k), k = 1, ...\}$ donde $\{\Theta_k\}$ son vectores aleatorios idénticos distribuídos independientementes, y cada árbol vota para la clase más popular del vector de entrada x.

2.3.2. Convergencia del método

Dado un conjunto de clasificadores $h_1(x), h_2(x), ..., h_k(x)$, y el conjunto de entrenamiento dibujado al azar de la distribución del vector aleatorio X e Y, se define la función margen como:

$$mg(X, Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j)$$

donde $I(.)$ es la función del indicador. El margen mide el grado en que el número promedio de votos en X e Y para la clase correcta, excede el voto promedio para cualquier otra clase. Cuanto mayor sea el margen, mayor será la confianza en la clasificación. El error

de generalización es:

$$PE^* = P_{X,Y}(mg(X,Y) < 0)$$

donde los subíndices X e Y indican la probabilidad sobre el espacio X, Y.

En *Random Forest*, $h_k(X) = h(X, \Theta_k)$.

Para un gran número de árboles, se deriva la *Ley Fuerte de Grandes Números* que dice: *A medida que aumenta el número de árboles, para todas las secuencias* $1, ..., PE^*$ *converge a:*

$$P_{X,Y}(P_\Theta(h(X,\Theta) = Y) - \max_{j \neq Y} P_\Theta(h(X,\Theta) = j) < 0)$$

Esto explica el motivo por el cual los bosques aleatorios no sobre-estiman a medida que se agregan más árboles, pero producen un valor que tiende al error de generalización.

2.3.3. Fuerza y correlación

El objetivo de estos métodos es inyectar al algoritmo la aleatoriedad justa para maximizar la independencia de los árboles manteniendo una precisión razonable. En el caso de los *Random Forest*, estas cualidades se miden a nivel del conjunto y se denotan como *fuerza y correlación*.

Para los bosques aleatorios, se puede derivar un límite superior para el error de generalización en términos de dos parámetros:

- medidas de la precisión de los clasificadores individuales
- dependencia entre ellos.

La interacción entre estos dos proporciona la base para entender el funcionamiento de los bosques aleatorios.

La fuerza del conjunto de clasicadores se define como:

$$s = E_{X,Y} mr(X,Y)$$

Un límite superior para el error de generalización es dado por:

$$PE^* \leq \bar{\rho}(1 - s^2)/s^2$$

Los dos ingredientes implicados en el error de generalización para los bosques aleatorios son la fuerza de los clasificadores individuales en el bosque, y la correlación entre ellos en términos de las funciones del margen crudo.

Hay simplificaciones en la situación de dos clases. La función margen es:

$$mr(X, Y) = 2P_{\Theta}(h(X, \Theta) = Y) - 1$$

El requisito de la fuerza es positivo. La función margen crudo es:

$$2I(h(X, \Theta) = Y) - 1$$

Y la correlación $\bar{\rho}$ es entre $I(h(X, \Theta) = Y)$ e $I(h(X, \Theta') = Y)$. En particular, si Y toma los valores $+1$ y -1 queda:

$$\bar{\rho} = E_{\Theta,\Theta'}[\rho(h(., \Theta), h(., \Theta')]$$

De modo que $\bar{\rho}$ es la correlación entre dos miembros diferentes del bosque promediada sobre la distribución Θ, Θ'.

El concepto de margen en los conjuntos de clasificadores sirve para medir la seguridad con la que el conjunto acierta o se equivoca en su predicción ya que es la diferencia entre la proporción de árboles que aciertan y los que se equivocan. Cuanto mayor sea el margen medio obtenido por un *Random Forest* sobre un conjunto de datos, mayor será su fuerza.

2.3.4. Usando características aleatorias

Algunos bosques aleatorios tienen un error de generalización consistentemente menor que otros. Para mejorar la precisión, la aleatoriedad inyectada tiene que minimizar la correlación $\bar{\rho}$ mientras se mantiene la resistencia. Los bosques aquí estudiados consisten en el uso de entradas seleccionadas al azar o combinaciones de entradas en cada nodo para cultivar cada árbol.

Estos nuevos métodos, de resultados muy similares, se denominan *Forest-RI* y *Forest-RC*. Ambos, se basan en la combinación de *bagging* y *árboles de decisión no podados* en los que las decisiones en cada nodo se toman considerando sólo un subconjunto de variables seleccionadas aleatoriamente.

¿Cuántas características seleccionar en cada nodo? Necesitamos de las estimaciones internas del error de generalización, la fuerza del clasificador y la dependencia entre ellos.

2.3.5. Uso de estimaciones *out-of-bag* para monitorear el error, la fuerza y la correlación

Supongamos un método para construir un clasificador de cualquier conjunto de entrenamiento. Dado un conjunto de entrenamiento específico T, formamos los conjuntos de entrenamiento de *boostrap* T_k, construimos los clasificadores $h(x, T_k)$ y dejamos que éstos voten para formar el *bagged predictor*. Para cada y, x en el conjunto de entrenamiento, se agrega los votos solamente sobre aquellos clasificadores para los cuales T_k no contiene y, x. A ésto se lo denomina clasificador fuera de bolsa. Entonces, la estimación fuera de bolsa para el error de generalización es la tasa de error del clasificador fuera de bolsa en el conjunto de entrenamiento.

Dado que la tasa de error disminuye a medida que aumenta el número de combinaciones, las estimaciones fuera de bolsa tienden a sobrestimar la tasa de error actual. Para obtener estimaciones imparciales fuera de bolsa, es necesario pasar más allá del punto en el que

converge el error del conjunto de pruebas. Sin embargo, a diferencia de la *validación cruzada*, donde el sesgo está presente, pero su extensión desconocida, las estimaciones fuera de bolsa son imparciales.

La fuerza y la correlación también se pueden estimar utilizando métodos fuera de bolsa. Esto da estimaciones internas que son útiles para entender la exactitud de la clasificación y cómo mejorarla.

A continuación se detallan dos métodos de *Random Forest*:

- **Forest-RI**: *Random Forest* usando selección de entrada aleatoria (características). Se forma seleccionando al azar, en cada nodo, un pequeño grupo de variables de entrada para dividir. El árbol crece utilizando la metodología *Classification and Regression Tree* (CART) hasta el tamaño máximo y no podar.
- **Forest-RC**: *Random Forest* utilizando combinaciones lineales de entradas. Si existen pocas entradas, digamos M, se definen más características tomando combinaciones lineales aleatorias de una serie de variables de entrada. Es decir, se genera una característica L, como el número de variables que se van a combinar. En un nodo dado, L variables son seleccionadas al azar y se suman junto con coeficientes que son números aleatorios uniformes en [-1, 1]. F combinaciones lineales son generadas, y luego se realiza una búsqueda sobre éstas para la mejor división.

2.3.6. Exploración del mecanismo de *Random Forest*

En algunas aplicaciones, como por ejemplo el análisis de experimentos médicos, es crítico entender la interacción de las variables que está proporcionando la exactitud predictiva. Un inicio en este problema se realiza mediante el uso de estimaciones internas fuera de bolsa y la verificación por repeticiones utilizando sólo variables seleccionadas.

Supongamos que existen M variables de entrada. Después de que cada árbol es construido, los valores de la m-ésima variable en los ejemplos fuera de bolsa son permutados aleatoriamente y los datos fuera de bolsa se ejecutan en el árbol correspondiente. La

clasificación dada para cada x_n que está fuera de bolsa se guarda. Esto se repite para $m = 1, 2, ..., M$. Al final de la carrera, la pluralidad de votos de clase fuera de bolsa para x_n con la m-ésima variable aumentada se compara con la etiqueta de clase verdadera de x_n para dar una tasa de clasificación errónea.

2.4. Random Forest Regression

Los bosques aleatorios para la regresión se forman creciendo árboles dependiendo de un vector aleatorio *p-dimensional* $X = (X_1, ..., X_p)^T$ que representa las variables de entrada, y una variable aleatoria Y que representa la respuesta de valor real. Los valores de salida son numéricos y se asume que el conjunto de entrenamiento se extrae independientemente de la distribución del vector aleatorio Y, X. El objetivo es encontrar una función de predicción $f(X)$ para predecir Y. La función de predicción está determinada por una *función de pérdida* $L(Y, f(X))$ definida para minimizar el valor esperado de la pérdida

$$E_{X,Y}(L(Y, f(X)))$$

Donde los subíndices denotan expectativas con respecto a la distribución conjunta de X e Y.

$L(Y, f(X))$ es una medida de la proximidad $f(X)$ a Y. Para *Random Forest Regressión*, L es el error de generalización cuadrático medio definido por:

$$E_{X,Y}(Y - f(X))^2$$

La predicción del bosque es el promedio de las predicciones de sus árboles:

$$f(x) = \frac{1}{K} \sum_{k=1}^{K} h_k(x)$$

Donde K es el número de árboles en el bosque.

Para un bosque, la predicción es el promedio de los términos de sesgo más la contribución promedio de cada característica.

El predictor del bosque aleatorio se forma tomando el promedio sobre k de los árboles $h(X, \Theta_k)$.

Para *Random Forest* regression se cumple lo siguiente:
Como el número de árboles en el bosque tiende a infinito, se verifica :

$$E_{X,Y}(I - av_k h(X, \Theta_k))^2 \to E_{X,Y}(I - E_\Theta h(X, \Theta))^2 \tag{2.4.1}$$

El lado derecho de (2.4.1) se denota como PE^* (forest) representa el error de generalización del bosque. Se define el error de generalización medio de un árbol como:

$$PE^*(tree) = E_\Theta E_{X,Y}(Y - h(X, \Theta))^2$$

Supongamos que para todo Θ, $EY = E_X h(X, \Theta)$. Entonces:

$$PE^*(forest) \leq \bar{\rho} PE^*(tree) \tag{2.4.2}$$

Donde $\bar{\rho}$ es la correlación ponderada entre los residuos $Y - h(X, \Theta)$ y $Y - h(X, \Theta')$ donde Θ, Θ' son independientes.

La fórmula (2.4.1) señala los requisitos para los *Random Forest Regression* precisos, baja correlación entre los residuos y los árboles con errores bajos. El bosque aleatorio disminuye el error promedio de los árboles empleados por el factor $\bar{\rho}$. La aleatorización empleada debe apuntar a una baja correlación.

2.4.1. Resultados empíricos en la regresión

En los *Random Forest Regression* usamos la selección de características al azar por encima del *bagging*. Por lo tanto, podemos utilizar el monitoreo proporcionado por la estimación

fuera de bolsa para dar estimaciones de $PE^*(forest)$, $PE^*(tree)$ y $\bar{\rho}$. Estos se derivan de manera similar a las estimaciones en la clasificación. Se utilizan las características formadas por una suma lineal aleatoria de dos entradas. Cuanto más características se utilicen, menor $PE^*(tree)$ pero mayor $\bar{\rho}$.

Una diferencia interesante entre la regresión y la clasificación es que la correlación aumenta muy lentamente a medida que aumenta el número de características utilizadas. El efecto principal es la disminución de $PE^*(tree)$. Por lo tanto, se requiere un número relativamente grande de características para reducir $PE^*(tree)$ y obtener un error de *test* óptimo.

2.4.2. Comentarios finales

Los resultados resultan ser insensibles al número de características seleccionadas para dividir cada nodo. Por lo general, al seleccionar una o dos características se obtienen resultados óptimos.

Los bosques aleatorios son eficaces para eliminar el ruido en los datos de entrada del modelo. Dada una larga lista de variables de entrada y un conjunto de datos potencialmente escaso, es muy probable que cualquier modelo predictivo descubra relaciones falsas entre esos insumos y la variable objetivo elegida. Esto se traduce en *overfitting* y el modelo no generaliza lo suficientemente bien como para el futuro de entrada que no ha visto.

Debido a que *Random Forest* construye muchos árboles usando un subconjunto de las variables de entrada disponibles y sus valores, contiene árboles de decisión subyacentes que omitieron la variable/característica generadora de ruido. Al final, cuando es el momento de generar una predicción se produce un voto entre todos los árboles subyacentes y gana el valor de predicción mayoritaria.

Ventajas:

- Maneja predictores categóricos de forma natural.
- Construye de forma rápida, incluso para grandes problemas.

- No hay supuestos de distribución formales.
- Se ajusta automáticamente a iteraciones no lineales.
- Selección automática de variables.

Desventajas:

- No es fácil de interpretar si el árbol es pequeño.
- Los nodos terminales no sugieren un agrupamiento natural.
- La imagen no brinda una idea de qué variables son importantes y dónde.

Precisión: Es competitivo con los métodos de aprendizaje de máquinas más conocidos.

Inestabilidad: Si los datos varian un poco, los árboles individuales cambian, pero el bosque es más estable porque es una combinación de muchos árboles [11][12].

Auscultación de pavimentos

3.1. Introducción

Durante el período de vida de los tramos de una ruta, se iniciará un proceso de deterioro tal que al final de su vida útil manifestará un conjunto de fallas que reducirán la calidad de circulación incrementando los costos de mantenimiento y de los usuarios (ver figura 3.1).

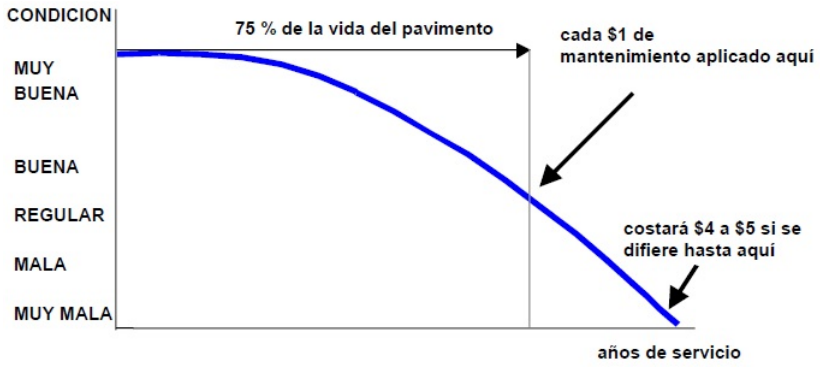


Figura 3.1: Evolución del comportamiento con los años de servicio y costos de mantenimiento asociados.

Para el ingeniero, el pavimento es una estructura formada por una o más capas, sobre la que actúan cargas en superficie, capaces de transmitir durante su vida útil las tensiones en profundidad. Para los usuarios, es una superficie que debe permitir la circulación del tránsito mixto, en condiciones de seguridad y comodidad, bajo cualquier condición climática, durante un tiempo prolongado.

Principalmente influenciado por dos factores: el clima, y el tránsito que debe soportar; será un buen diseño el que, con un costo de transporte anual mínimo, tenga en cuenta

simultáneamente ambos factores, en la medida de su importancia. Una vez puesto en servicio, el pavimento comienza a tener respuestas favorables o desfavorables que deben evaluarse permanentemente.

Independientemente del deterioro natural, se deben realizar ciertas obras de mantenimiento y rehabilitación en tiempo oportuno, con el objeto de reducir el impacto que las diferentes fallas pueden afectar a la estructura, a fin de optimizar los recursos disponibles para evitar trabajos de reconstrucción de la estructura. Para esto se requiere de técnicas de gestión de pavimentos. Además de la auscultación de calzadas pavimentadas, que consiste en la valoración del estado de los deterioros superficiales y estructurales que afectan la calidad del servicio brindado a los usuarios; utilizando metodologías de análisis definidas y equipamiento apropiado [13].

3.2. Definición

Existen dos tipos de evaluación de pavimentos. La evaluación superficial y la estructural. La primera, traducida en parámetros como textura, rugosidad, fricción, señalización, etc.; evalúan la calidad, seguridad y comodidad de circulación. A pesar de estar asociada con su capacidad estructural, pueden existir sectores con buena calidad de circulación y sin embargo poseer una estructura débil, que con el paso de la carga se fatigará y presentará un agrietamiento severo. El segundo tipo de evaluación se asocia al control de calidad estructural, mide distintos parámetros de fallas en la ruta, tales como la deflexión, espesor, densidad, humedad.

Con la recolección de los datos de las diferentes evaluaciones del estado de la carretera a través de los años, utilizando metodologías de análisis definidas y equipamiento apropiado, se puede priorizar obras de mantenimiento y diseño de mejoras.

Se definen umbrales de estado de cada parámetro, para proponer mantenimiento de rutina si el parámetro es aceptable, o mejoras si supera el umbral superior (ver figura 3.2) [13][1].

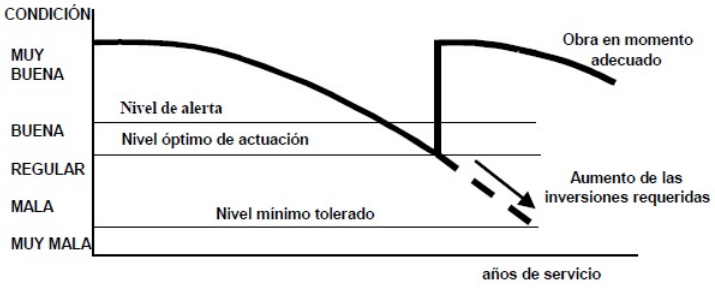


Figura 3.2: Niveles de condición.

3.3. Indice de Rugosidad Internacional

Existen diferentes parámetros (deterioros) para la valoración del comportamiento de la superficie de la carretera: rugosidad, ahuellamiento, fisuras, baches, desprendimientos, exudación, etc.

La característica que se pretende evaluar para el estado de deterioro del pavimento es la rugosidad, que valora las desviaciones del perfil longitudinal del camino, respecto a una superficie plana. La misma, afecta en forma muy importante la dinámica del vehículo, la calidad de circulación, el efecto dinámico de las cargas y el drenaje. Los defectos de rugosidad son percibidos por los usuarios como movimientos vibratorios que afectan el confort de circulación.

El indicador que representa a este parámetro es el IRI (Indice de Rugosidad Internacional, definido en 1982 por el Banco Mundial), el cual es un índice obtenido mediante una simulación matemática del pasaje de un vehículo virtual, circulando sobre el perfil del camino a una velocidad de 80 km/h.

La rugosidad IRI es una variable de tipo continua; donde el rango de la misma define una escala que va desde 0 para superficies idealmente planas (las pistas de aeropuertos pueden brindar valores inferiores a la unidad), tomando valores entre 1 y 2 m/km para pavimentos recién construídos, y valores superiores a 6 m/km para superficies notoriamente deterioradas (ver figura 3.3)[14][13].

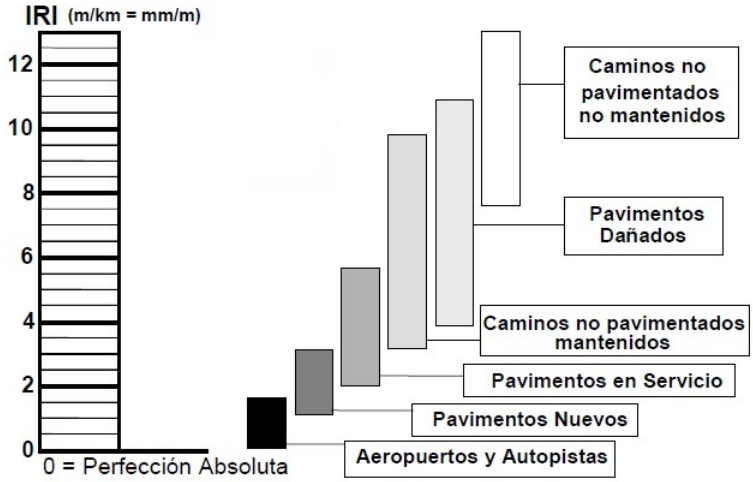


Figura 3.3: Escala de rugosidad IRI.

3.3.1. Formas de medición

Existe una gran variedad de equipos y formas de medición, pero todas deben expresar sus resultados en una misma unidad de medición IRI(m/km). Estos equipos miden la falta de confort en una unidad propia de cada uno. Para que los resultados puedan ser comparados con otros equipos deben ser correlacionados al Índice Internacional de Rugosidad IRI tomado como referencia. Se seleccionan tramos de caminos en servicio. Se mide mínimo 5 veces con el equipo a la velocidad de operación; sobre la huella externa.

Si se detecta la notoria preponderancia de un carril respecto del otro, la medición con el equipo de relevamiento se realizará sobre esta trocha.

3.3.2. Errores asociados a los equipos

Error de estimación del IRI, es la diferencia entre el valor informado por el equipo y el valor real de la superficie. Este error contempla tanto la repetibilidad de la medición como el error asociado con la calibración al valor IRI.

En los equipos homologados para realizar mediciones de control en tramos, el límite de error de tolerancia de medición en los equipos de rugosidad es de 0.3 m/km.

Aunque el error sea reducido, para penalización deben realizarse tres mediciones y analizar el promedio.

Durante la medición en tramos están presentes otros errores que generan diferencias entre pasadas repetidas, como así también diferencias entre equipos:

- Línea de medición: una variación de algunos centímetros entre líneas de medición origina variaciones en los resultados, esto es despreciable en carreteras nuevas, pero puede ser crítico en carreteras en servicio con deformaciones no homogénea.
- Velocidad de medición: si bien la misma está limitada a +/- 5 km/h respecto a la velocidad de operación del equipo, este valor se muestra en los informes como promedio en tramos de 100 metros y puede tener variaciones dentro del mismo que solo son controladas por el chofer del equipo durante la medición.
- Aceleraciones y desaceleraciones bruscas en el tramo, que alteran el funcionamiento del sistema dinámico.

La huella externa de cada carril es la más deteriorada desde el punto de vista de las deformaciones que afectan a la rugosidad, ya que el borde externo es una zona donde la calidad de la construcción es más deficitaria, favoreciendo el ingreso de agua y provocando una reducción de la capacidad estructural [1].

Modelado de la red vial de Entre Ríos

El relevamiento periódico de la condición del pavimento en forma ordenada y sistemática, así como también la evolución de los deterioros del mismo, permiten conformar la función del comportamiento de los distintos tramos de una ruta, para poder predecir las tareas de mantenimiento necesarias en magnitud y oportunidad.

4.1. Selección de variables independientes

Las variables independientes, predictoras en lo sucesivo, son independientes en el sentido de que entran en el modelo desde fuera, son variables externas y mesurables.

La elección de estas variables es, no menos importante que la elección de la variable objetivo; ya que determina el éxito del modelado. La mayor parte del tiempo invertido en el desarrollo de modelos se emplea precisamente en el análisis y la elección del conjunto de variables independientes.

Se parte de un modelo donde las características empleadas son:

- identificación del tramo
- año de la medición
- deflexión
- TMDA

El *tránsito* es la variable más importante en el diseño de las rutas, ya que se estudia los efectos que las cargas de los vehículos causarán sobre las mismas. Por esto, se debe

conocer el número y tipo que circulará por una vía, así como la intensidad de la carga y la configuración del eje que la aplica.
Como información inicial, se tiene el dato del TMDA, correspondiente al *tránsito medio diario anual*, que se define como el volumen del tránsito total anual, dividido por el número de días del año. Corresponde al análisis de los espectros de carga por tipo de eje, a partir de información detallada sobre el uso de las vías en períodos de horas, días, semanas, meses y años.

Se define como *deflexión* a la deformabilidad elástica de la estructura que se produce bajo la carga, evaluada desde la superficie. Es la respuesta de los pavimentos ante un estímulo, en general cargas impuestas por el tráfico. Es la deformación que se registra en la superficie del pavimento cuando es sometido a cargas. En definitiva la deflexión permite estudiar la capacidad estructural existente de un pavimento [13].

4.2. Tipo de información recolectada

En el Laboratorio Vial del Instituto de Mecánica Aplicada y Estructuras, Facultad de Ciencias Exactas, Ingeniería y Agrimensura de la Universidad Nacional de Rosario se desarrolló un Sistema de Administración Vial para su implementación en la Dirección Provincial de Entre Ríos. A través de dicho convenio se dispuso de los datos relevados de dicha provincia.

La base de datos guarda información relacionada con la historia constructiva de la ruta (fechas, materiales), historia de los mantenimientos realizados, tránsito, modelos estructurales obtenidos del análisis de la misma. Como se mencionó anteriormente, la descripción actual del estado de la red, se realiza a lo largo de los distintos *tramos homogéneos* en que ésta se subdivide en la totalidad de la red. Estos tramos homogéneos constituyen la unidad de análisis de la información y se definen como los sectores que presentan características iguales de estructura y tránsito. Estos tramos son independientes unos de otros.

Se dispone de resultados de mediciones periódicas de la rugosidad medidos desde el año 2009 al 2015. Habiendo recolectado la información disponible de un sistema de gestión y administración, se disponen de 189 conjuntos de datos, correspondientes a 27 tramos homogéneos diferentes.

La tabla 4.1 muestra una pequeña parte de los datos de Entre Ríos.

Tramo	Año de medición	Deflexión	TMDA	IRI
1	2009	41	1051	2.403
1	2010	41	1051	2.538
1	2011	41	1051	2.389
1	2012	43	1346	2.386
1	2013	43	1162	2.298
1	2014	43	1162	2.251
1	2015	43	1162	2.063
2	2009	77	1051	2.741
2	2010	77	1051	2.863
⋮	⋮	⋮	⋮	⋮
2	2015	75	1162	2.224
⋮	⋮	⋮	⋮	⋮
27	2009	70	1200	2.056
27	2010	70	1200	2.233
27	2011	70	1200	2.359
27	2012	70	1200	2.570
27	2013	70	752	1.920
27	2014	70	752	2.218
27	2015	70	752	2.132

Tabla 4.1: Datos recolectados de la Red Vial de Entre Ríos.

La calidad del paquete estructural y el terreno, valorados en su conjunto por la deformabilidad bajo carga (deflexión) es un dato que se asume casi constante a través de los años en los casos que no fue reevaluado.

El TMDA no siempre tiende a crecer, como se observa en la tabla 4.1, en el tramo 1, del año 2012 al 2013 decae de 1346 a 1162. Este dato se ajusta a la economía del país, más vehículos se pueden comprar, más tránsito circula por la ruta, pero si hay una crisis, el tránsito decae notoriamente. También puede ocurrir que se inaugure otro camino que haga el mismo tramo, de esta manera el tránsito se divide entre ambos caminos.

Y en la última columna se tiene la rugosidad, medida anualmente.

Las diferentes entradas del tramo 1, caracterizado en la tabla 4.1, significa que a través de los años consecutivos desde el 2009 al 2015, además de las características tramo, deflexión,

y TMDA; se dispone de la variable objetivo, la *rugosidad*.

4.3. Selección del conjunto entrenamiento/validación

Como selección del conjunto de entrenamiento y validación se pensó en tres selecciones.

1. Como primera selección del conjunto se intentará predecir la rugosidad para el último año conocido. Dicho de otra manera, se considera como conjunto de entrenamiento los primeros $n-1$ años conocidos de cada tramo, dejando el último año para validación del modelo. Conformando de esta manera 162 conjuntos de entrenamiento y 27 de validación.

   Como el tamaño del conjunto es pequeño, se podría pensar en el uso de la *validación cruzada* que es una especie de entrenamiento/validación repetido en el que las particiones de *test* nunca solapan. Se divide una vez el conjunto de datos en varios grupos y se "mezcla" cual es de entrenamiento y cual de prueba, y se calcula la media.

   En la validación cruzada de K iteraciones o *K-fold Cross-Validation* los datos se dividen en K subconjuntos (*folds*). Uno de los subconjuntos se utiliza como datos de prueba y el resto $(K-1)$ como datos de entrenamiento. El proceso de validación cruzada es repetido durante K iteraciones, con cada uno de los posibles subconjuntos de datos de prueba.

   En este análisis del modelo se aplicaría un *K-fold Cross-Validation* con $K = \#(tramos)$, pero esto no es correcto puesto que los 27 tramos que se pretende predecir corresponde a un año específico, es decir, al último año medido. No es aleatorio la selección de los 27 conjuntos de validación. Por tal motivo, no se hace uso de la *validación cruzada*.

2. Como variante del conjunto de entrenamiento, se plantea la posibilidad de entrenar con todos los tramos menos uno, y validar el modelo con el tramo restante. Esto

es lo que se denomina *Leave-one-out*, que es una validación cruzada con $k =$ *espacio de datos de entrenamiento*.

Es decir, para N datos de entrenamiento, repetir $K = N$ veces:

- Reservar el dato número N para *test*.
- Entrenar con los $N - 1$ datos restantes.
- Hacer el testeo con el dato N.

Es preciso porque se usan casi todos los datos para entrenar, y a la vez todos los datos figuran como *test* en alguno de los ciclos. Pero es costoso en tiempo (hay que lanzar el algoritmo de aprendizaje N veces).

Ahora bien, ¿tiene sentido predecir la rugosidad de todo un tramo?
Si existe el tramo, se debiera contar con los datos iniciales del proyecto de dicha carretera. Lo que se pretende predecir es el estado de deterioro de circulación para un año determinado. Más precisamente para un año mayor al último año conocido. El análisis con *Leave-one-out* no tiene sentido en la vida útil de un tramo.

3. Por último, y como mencionamos anteriormente se intentará predecir la rugosidad para un año determinado, superior al último conocido. Por lo que como análisis final, se entrenará con los datos de validación. Como el *tiempo* no es algo natural de la regresión, se generará un vector de entrada para testear de la siguiente manera: el año a predecir será el año acumulado incrementándose de a una unidad. Tanto para la deflexión como para el tránsito liviano, medio y pesado, se le calculará un incremento del 2 % anual. Se necesitará pasar la rugosidad del año anterior como *input*. Se agregará la columna con el IRI del año anterior, prediciendo año tras año (hasta llegar a los años que se pretende predecir, es decir qué sucede para el tercer año después de la última medición), testeando con el IRI predicho.

4.4. Evolución del modelo experimental

En esta sección se describirá los diferentes análisis progresivos que se fueron realizando, usando como base los conjuntos de entrenamiento y validación descriptos anteriormente. Para el desarrollo de las técnicas de prueba se utilizó como lenjuage a *Python*. Y como paquetes de software que incluyen algoritmos de aprendizaje automatizado, a *Scikit-learn*. Esta librería también facilita las tareas de evaluación y diagnóstico ya que proporciona varios métodos de fábrica para poder realizar estas tareas en forma más simple.

Muchos elementos utilizados en la función objetivo de un algoritmo de aprendizaje (como el núcleo kernel de base radial de *Vector Support Machine*) asumen que todas las características están centradas alrededor del cero y tienen varianza en el mismo orden. Si existe una característica que tiene órdenes de magnitud mucho mayores al resto puede dominar totalmente el aprendizaje. Podría dominar la función objetivo y hacer que el estimador sea incapaz de aprender de otras características correctamente como se esperaba.

La *normalización* de los conjuntos de datos es un requisito común para muchos estimadores de aprendizaje automático implementado en *Scikit-learn*; que podrían comportarse mal si las características individuales no se parecen más o menos a los datos estándar normalmente distribuidos: gaussiana con media cero y varianza unitaria.
A menudo en la práctica, se ignora la forma de la distribución y simplemente se transforman los datos para centrarlos, eliminando el valor medio de cada característica y luego escalándolos dividiendo rasgos no constantes por su desviación estándar.

En *Python*, se cuenta con la biblioteca *preprocessing* que proporciona una forma rápida y fácil de realizar esta operación en un solo conjunto de datos [15].

Para entrenar con *Support Vector Machine Regression* se normalizó el vector de entrada sobre todos los datos (menos la rugosidad), es decir las características se reajustaron de modo que tengan las propiedades de una distribución normal estándar con $\mu = 0$ y $\sigma = 1$.

Tanto para *Support Vector Machine Regression* como *Random Forest Regresion* se opti-

mizaron los parámetros de los modelos de regresión con la utilización del método *Grid Search*, que consiste en un barrido de parámetros en búsqueda del mejor:

- *Support Vector Machine Regression*:
  - *constante de regularización* de 1e-3, 0.01, 0.1, 1, 10. Luego barre dentro del que encontró (por ejemplo si encuentra el 0.1 barre en 0.1 0.3 0.5 0.7 0.9).
  - *épsilon* de 1e-3, 0.01, 0.1, 1, 10.
- *Random Forest Regresion*:
  - *cantidad de árboles* de 100, 200, 300, 400, 500, 600, 700, 800
  - *profundidad* de 4, 5, 6, 7.

Para ambos modelos el error que se consideró es el *error cuadrático medio* (mse).

4.4.1. Entrenando con los primeros $n-1$ años

En las figuras empleadas en esta primer selección de conjunto de entrenamiento/validación se grafican el valor real de la rugosidad correspondiente al último año medido, junto con el valor predicho por cada modelo. En el eje x se indican los diferentes tramos que fueron evaluados, y en la leyenda se complementa la información con el error cuadrático medio de cada modelo.

La figura 4.1 muestra ambos modelos de regresión para el primer análisis, con los datos originales de la tabla 4.1.

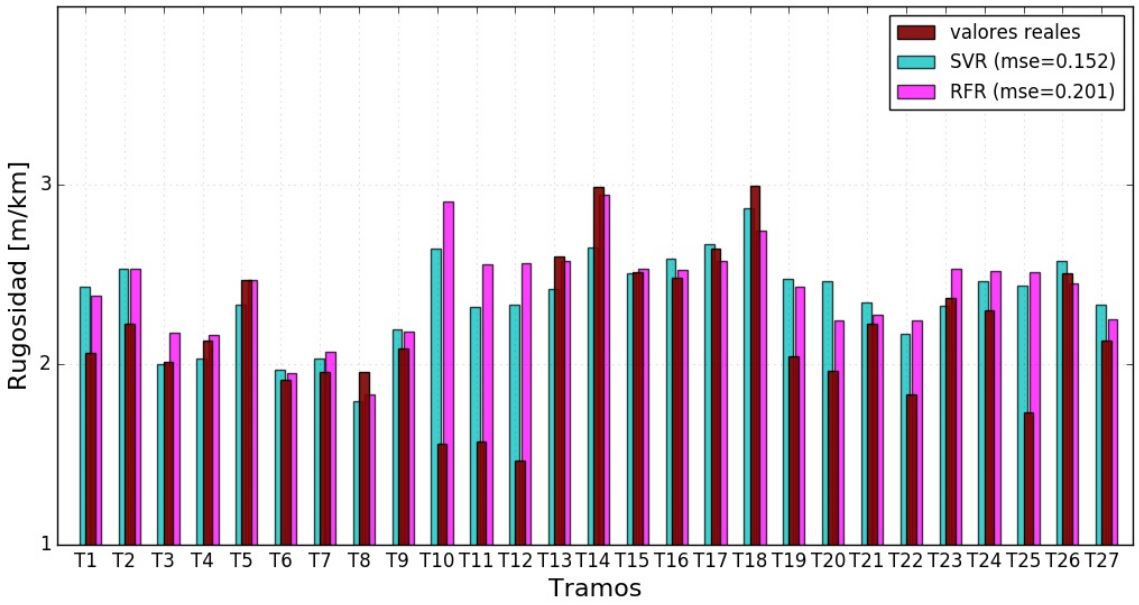


Figura 4.1: Modelos de regresión para los datos originales.

En la tabla 4.2 se visualiza los errores cuadráticos de los mismos con la selección de los mejores parámetros.

Modelo	Parámetros	Error $[m/km]^2$
SVR	$C = 5$ - $\varepsilon = 0.1$	0.152
RFR	400 árboles - profundidad = 4	0.201

Tabla 4.2: Errores cuadráticos para la corrida con los datos originales.

Filtrando mejoras

En el figura 4.1, existen varios picos, en los que el regresor pareciera no comportarse de manera correcta. Se puede observar la diferencia notable de ambos modelos para los tramos 10, 11 y 12 por ejemplo. En la tabla 4.3 se visualizan los datos crudos de los 9, 10 y 11 tramos mencionados. Se tratará el tramo 9 como parte de la misma problemática.

A través de los años, la rugosidad tiende a crecer, lo que equivale a que la calidad empeora. Sin embargo, en el tramo 9, del año 2009 al 2010 la rugosidad baja notoriamente su valor de 2.522 a 1.771, dando una diferencia de 0.75 m/km IRI; lo que significa que en dicho tramo, en el año 2010 se realizaron *obras de mantenimiento o mejoras*. Lo mismo ocurre

Tramo	Año de medición	Deflexión	TMDA	IRI
9	2009	60	2070	2.522
9	2010	60	2070	1.771
9	2011	60	2070	2.190
9	2012	60	2070	2.386
9	2013	60	2070	2.086
9	2014	60	2070	2.147
9	2015	60	2070	2.091
10	2009	60	3319	2.438
10	2010	60	3319	2.778
10	2011	60	3319	2.620
10	2012	60	3319	2.846
10	2013	60	3319	3.053
10	2014	60	3319	3.746
10	2015	60	3319	1.558
11	2009	63	1684	2.246
11	2010	63	1684	2.470
11	2011	63	1935	2.255
11	2012	63	1935	2.399
11	2013	63	2092	2.570
11	2014	63	2092	2.456
11	2015	63	2092	1.569

Tabla 4.3: Datos de los tramos 9, 10 y 11.

en el año 2015 para los tramos 10 y 11. Se considera como mejora cuando la diferencia de la rugosidad con respecto al año anterior excede en 0.6 m/km IRI.

Se plantea una modificación al código del regresor donde se descarta años anteriores a la mejora si la cantidad de años siguientes medidos superan a los años anteriores a la misma. En caso contrario, se entrena con los años anteriores. Se hace esta diferencia para no descartar el tramo completo (como sería el caso del tramo 10).

La figura 4.2 muestra ambos modelos de regresión. Y en la tabla 4.4 se visualizan los errores cuadráticos de los mismos con los parámetros optimizados.

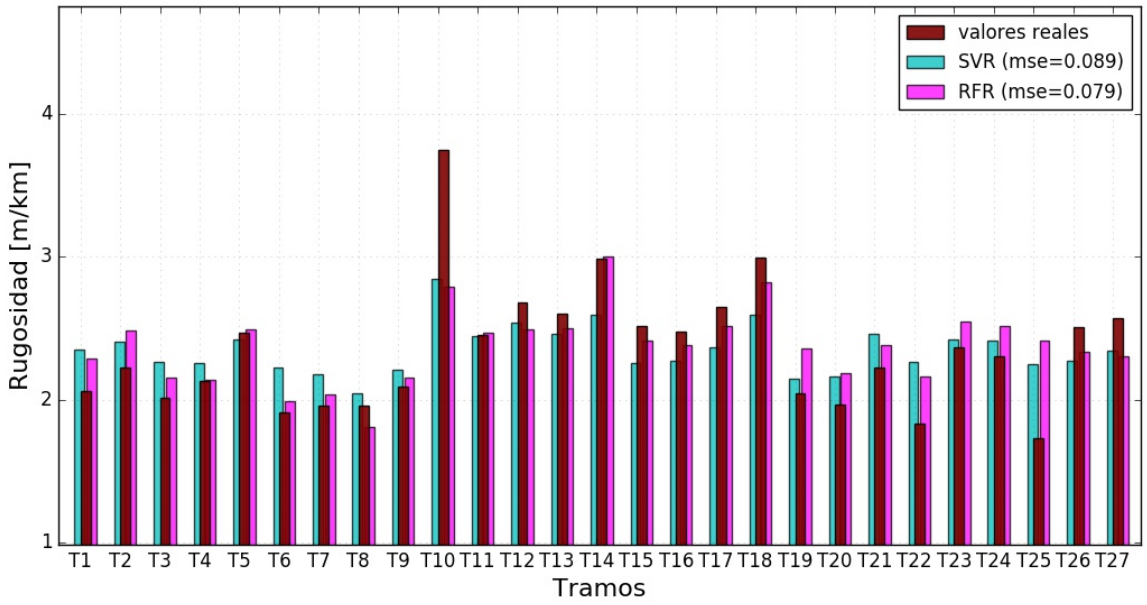


Figura 4.2: Modelos de regresión filtrando mejoras.

Modelo	Parámetros	Error $[m/km]^2$
SVR	$C = 1$ - $\varepsilon = 0.1$	0.089
RFR	700 árboles - profundidad = 7	0.079

Tabla 4.4: Errores cuadráticos con eliminación de mejoras.

Como se puede observar, mejoró notoriamente el regresor ya que se redujo el error pasando de $0.152(m/km)^2$ a $0.089(m/km)^2$ con *Vector Support Machine* y de $0.201(m/km)^2$ a $0.079(m/km)^2$ con *Random Forest*.

Descomposición de la variable independiente TMDA

¿Se podría mejorar el modelo si se mejoran las variables independientes?

El TMDA o tránsito medio diario anual se define como el volumen del tránsito total anual, dividido por el número de días del año. El tránsito es la variable más importante en el diseño de las rutas, ya que se estudia los efectos que las cargas de los vehículos causarán sobre las mismas. Por esto, se debe conocer el número y tipo que circulará por una vía, así como la intensidad de la carga y la configuración del eje que la aplica. No es lo mismo que el TMDA sea 400 y se componga de 100 camiones y 300 autos. Que sea en

su totalidad de 400 camiones. El efecto sobre la ruta será diferente.

Por tal motivo, el siguiente paso será descomponer el TMDA en tránsito liviano, mediano y pesado, que se corresponderá a automóviles, omnibus y camiones respectivamente.

- Automóviles: se incluyen también en esta categoría a las camionetas y todo otro vehículo cuyas características de operación se asemeje a las de los automóviles.
- Omnibus: incluye a los colectivos, micro-omnibus y similares.
- Camiones: incluye a los camiones con y sin acoplado, semi-remolques, semi-remolques con acoplado y todo otro vehículo cuyas características de operación sean similares a las de los camiones.

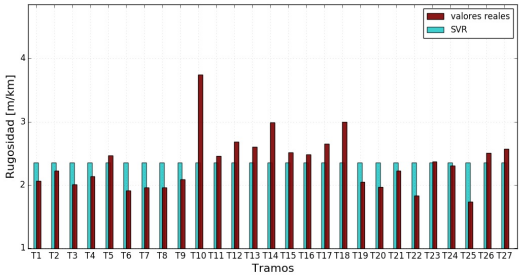
En la siguiente tabla se visualizan los datos del tramo 1 con el TMDA desglosado en la clasificación anterior.

Tramo	Año de medición	Deflexión	Tránsito Liviano	Tránsito Medio	Tránsito Pesado	IRI
1	2009	41	881	13	157	2.403
1	2010	41	881	13	157	2.538
1	2011	41	1057	13	266	2.389
1	2012	43	885	23	253	2.386
1	2013	43	885	24	253	2.298
1	2014	43	885	24	253	2.251
1	2015	43	885	24	253	2.063

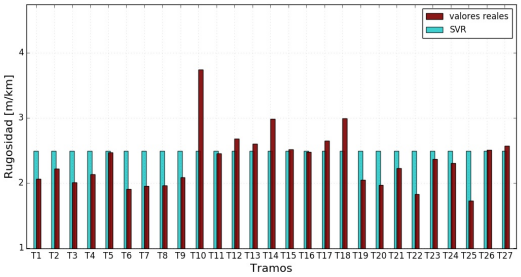
Tabla 4.5: Datos del tramo 1 con TMDA desglosado.

Ahora se disponen de 6 variables independientes, y la misma variable objetivo.

Cuando C tiende a un número pequeño, el resultado es casi constante. A medida que se incrementa, el gráfico va tomando curva "semejante" al modelo original, hasta que el error se minimiza y se hace constante.
Con ε, se va minimizando el error a medida que se incrementa su valor, hasta que el gráfico se hace constante como muestra la siguiente figura.



(a) Resultado constante con C=1e-3.



(b) Resultado constante con ε=1.

Figura 4.3: Modelos de regresión con TMDA desglosado para C=1e-3 y ε=1.

En la figura 4.4 se muestra ambos modelos de regresión. En la tabla 4.6 se visualizan los errores cuadráticos de los mismos, con los parámetros optimizados.

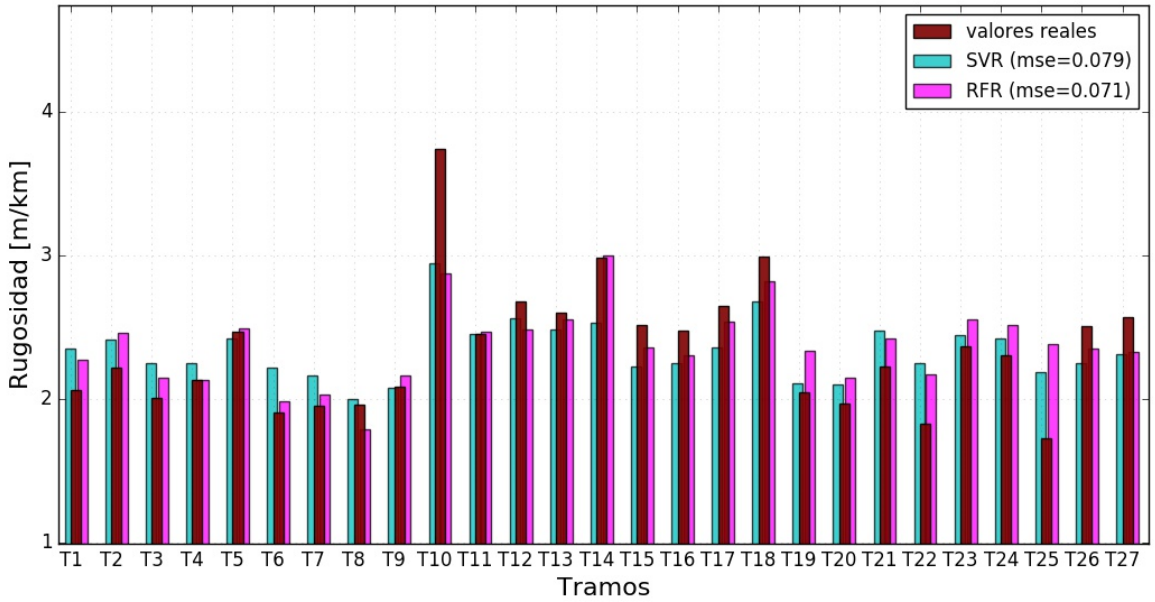


Figura 4.4: Modelos de regresión con TMDA desglosado.

Modelo	Parámetros	Error $[m/km]^2$
SVR	$C = 4$ - $\varepsilon = 0.1$	0.079
RFR	100 árboles	0.071

Tabla 4.6: Errores cuadráticos con TMDA desglosado.

Para este caso, el parámetro optimizado con *Random Forest Regression* no se definió la máxima profundidad de cada árbol, así, los nodos se expanden hasta que todas las hojas

tengan menos muestras que la variable *min_samples_split* (que es el número mínimo de muestras necesario para dividir un nodo interno, que por defecto está en 2).

Como se observa, el error es muy similar en ambos modelos y mejoró muy levemente respecto al anterior análisis.

Si se observa la evolución de la importancia de las características con *Random Forest Regression*, a través del comando *feature_importances_* se observa lo siguiente:

- Con el análisis de los datos originales el vector de importancia es: $[0.552, 0.066, 0.227, 0.155]$. Lo que determina que la característica más relevante es la identificación del tramo.
- Filtrando las mejoras, el vector de importancia está determinado por: $[0.520, 0.127, 0.222, 0.131]$. Se observa que la identificación del tramo está perdiendo peso, y va tomando impulso el año de la medición del IRI.
- Desglozando el TMDA, el vector fue: $[0.219, 0.125, 0.178, 0.089, 0.331, 0.058]$. La suma de las importancias discriminadas del TMDA (0.478) resulta mayor a la anterior en conjunto, es decir que aporta mucha información al modelo que se está intentando describir.

4.4.2. Entrenando con todos los tramos menos uno

Se analizará modificando los datos de entrenamiento y validación.

Si dos tramos tienen las mismas entradas, pero varía la rugosidad entre ambos, el regresor predice la misma salida y esto no es correcto. Esto ocurre porque no se puede predecir la evolución temporal de la salida, sin entradas que cambien con el tiempo. De esta manera, el siguiente paso será modificar las entradas, por los acumuladores a través de los años. La tabla 4.7 muestra cómo quedaría modificado un tramo por su acumulado.

Se repitió el procedimiento de optimización de parámetros.
Las figuras empleadas en esta segunda selección de conjunto de entrenamiento/validación

Tramo	Año	Deflexión	Tránsito Liviano	Tránsito Medio	Tránsito Pesado	IRI
1	0	41	881	13	157	2.403
1	1	82	1762	26	314	2.538
1	2	123	2819	39	580	2.389
1	3	166	3704	62	833	2.386
1	4	209	4589	86	1086	2.298
1	5	252	5474	110	1339	2.251
1	6	295	6359	134	1592	2.063

Tabla 4.7: Datos del tramo 1 con entradas acumuladas.

grafican para un tramo determinado, el valor real de la rugosidad correspondiente a cada año medido, junto con el valor predicho de cada año por cada modelo. En el eje x se indican los diferentes años acumulados que fueron evaluados, y en la leyenda se complementa la información con el error.

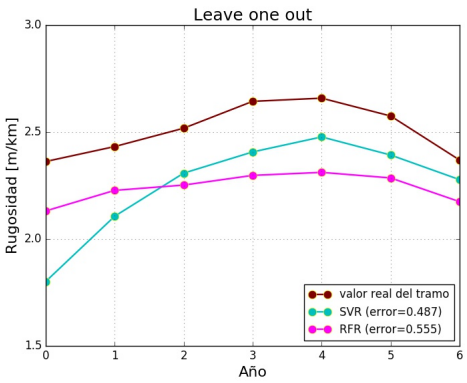
Pero ahora el error se calcula sobre la diferencia de área entre curvas. Es decir, sobre la superficie de cada área entre dos mediciones válidas. El error de un tramo se valora como la sumatoria de las diferencias por cada año evaluado dividido el número de años. Y para el error total, se promedia los errores correspondientes al conjunto de tramos.

En la tabla 4.8 se visualiza donde se dió el valor óptimo para cada método, con su respectivo error.

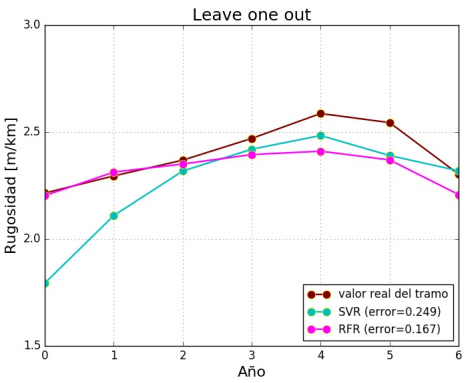
Modelo	Parámetros	Error $[m/km]$
SVR	$C = 4$ - $\varepsilon = 0.1$	0.559
RFR	800 árboles - profundida máxima=4	0.233

Tabla 4.8: *Leave-one-out* con acumulados.

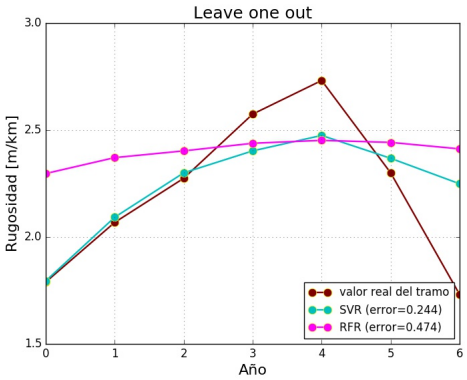
En la figura 4.5 se visualizan las curvas para algunos tramos.



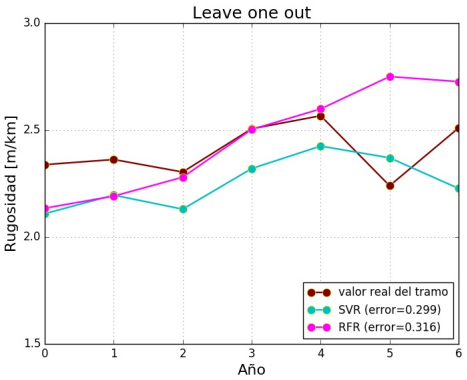
(a) Tramo 23.



(b) Tramo 24.



(c) Tramo 25.



(d) Tramo 26.

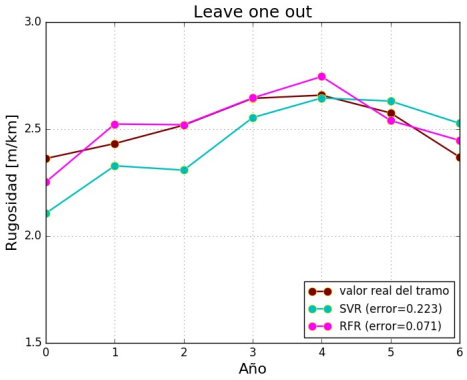
Figura 4.5: *Leave-one-out* con acumulados.

Manteniendo el IRI del año anterior

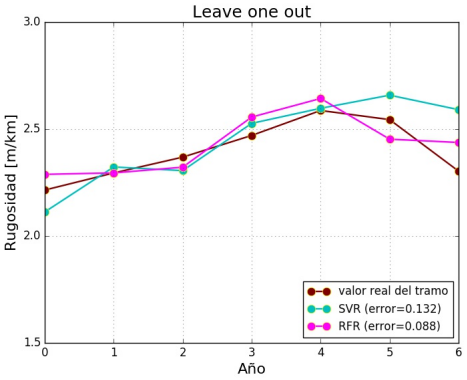
La siguiente modificación que se hará al modelo, será que en lugar de acumular los datos, se agregará una columna con el IRI del año anterior. La variable del tramo no la consideraremos como entrada, ya que es una variable externa con ruido innecesario. Solo se acumulará la variable del año. Para el IRI anterior del año inicial, se hace una interpolación lineal, y se le agregará ruido Gaussiano con desvío estándard de 0.5.

El error se consideró, al igual que en la anterior prueba, sobre la superficie de cada área entre dos mediciones válidas.

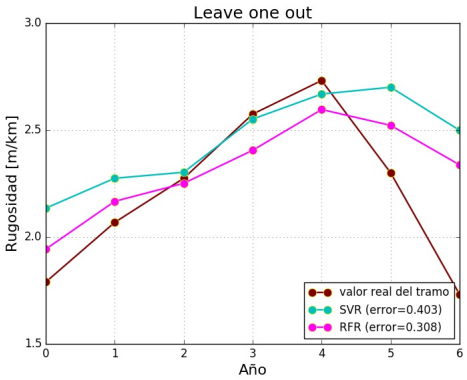
En la tabla 4.9 se visualizan los valores óptimos de cada método, con su respectivos parámetros.



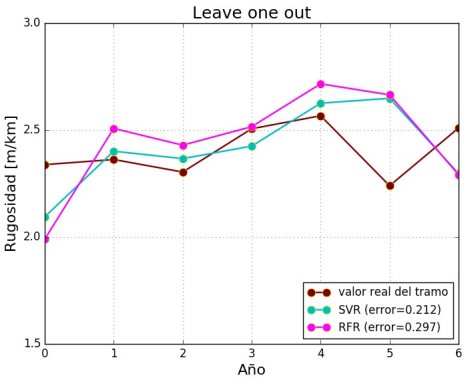
(a) Tramo 23.



(b) Tramo 24.



(c) Tramo 25.



(d) Tramo 26.

Figura 4.6: *Leave-one-out* con IRI del año anterior.

Modelo	Parámetors	Error $[m/km]$
SVR	$C = 0.1$ - $\varepsilon = 0.1$	0.543
RFR	600 árboles	0.147

Tabla 4.9: *Leave-one-out* con IRI del año anterior.

Al agregarle el IRI del año anterior, y eliminar el tramo para reducir el ruido, el modelo mejoró la salida. Si bien el resultado con *Support Vector Machine* resulta ser relativamente bueno, no resultó ser mejor al compararlo con *Random Forest*. Si se observa el tramo 23 de la figura 4.5 y 4.6 se puede visualizar como *Random Forest* predice los valores exactos para el segundo y tercer año. También se observa para los tramos presentados de esta última modificación del modelo, que las curvas de ambos se "asemejan" un poco más a la real.

4.4.3. Entrenando con los datos de validación

Pasando a la última selección de conjunto de entrenamiento y validación, se intentará predecir la rugosidad para un año determinado (extrapolar), superior al último. Como mencionamos anteriormente, el *tiempo* no es algo natural de la regresión, por lo que se generará un vector de entrada para testear de la siguiente manera: tramo va a ser el mismo, el año a predecir será el año acumulado incrementándose de a una unidad. La deflexión se mantiene constante, por lo que toma el valor del último medido. Para el

tránsito liviano, medio y pesado, se le calculará un incremento del 2 % anual. Se necesitará pasar la rugosidad del año anterior como *input*. Se agregará la columna con el IRI del año anterior, prediciendo año tras año (hasta llegar a los años que se pretende predecir, es decir qué sucede para el tercer año después de la última medición), testeando con el IRI predicho.

Como se mencionó en el capítulo 3 existen errores asociados a los equipos. Más específicamente, error de estimación del IRI, que es la diferencia entre el valor informado por el equipo y el valor real de la superficie. El límite de error de tolerancia de medición en los equipos de rugosidad es de 0.3 m/km. Una variación de algunos centímetros entre líneas de medición origina variaciones en los resultados, esto es despreciable en carreteras nuevas, pero puede ser crítico en carreteras en servicio con deformaciones no homogénea.

El siguiente paso será eliminar aquellas mediciones que se encuentren entre 0.3 m/km y 0.6 m/km (si supera este último parámetro significa que hubo obra de mejora). Para reducir el error, se aproximará por una polinómica de grado 3 para generar la rugosidad de esos años eliminados.

Se repitió el procedimiento de búsqueda de optimización de parámetros. Primero para predecir el IRI del último año conocido (como en el sección 4.4.1).

En la figura 4.7 se puede visualizar las curvas con ambos modelos. Y en la tabla 4.10 muestra los errores cuadráticos de los mismos con los parámetros optimizados.

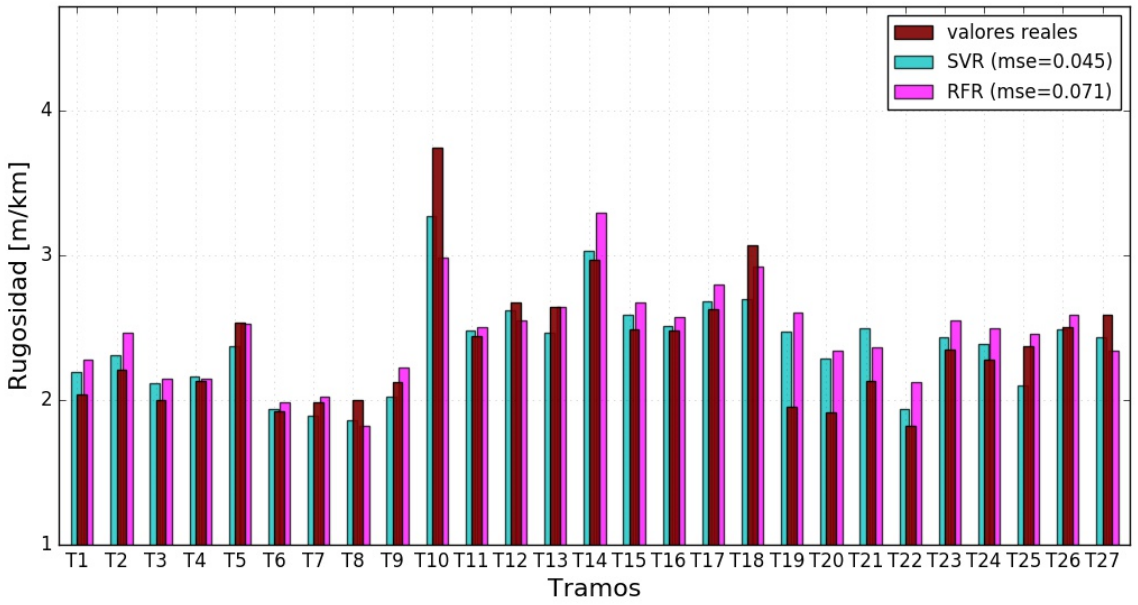


Figura 4.7: Predicción del último año conocido.

Modelo	Parámetros	Error $[m/km]^2$
SVR	$C = 5$ - $\varepsilon = 0.001$	0.045
RFR	800 árboles	0.071

Tabla 4.10: Errores para predicción del último año conocido.

Ahora se intentará predecir el IRI para un vector generado, donde la entrada para el *año* es superior en 3 al último año conocido. De otra manera, se intentará predecir la rugosidad a 3 años futuros, que es lo que una gestión política desearía conocer para realizar las obras necesarias de mantenimiento.

La figura 4.8 muestra el resultado de ambos modelos con los parámetros anteriormente optimizados.

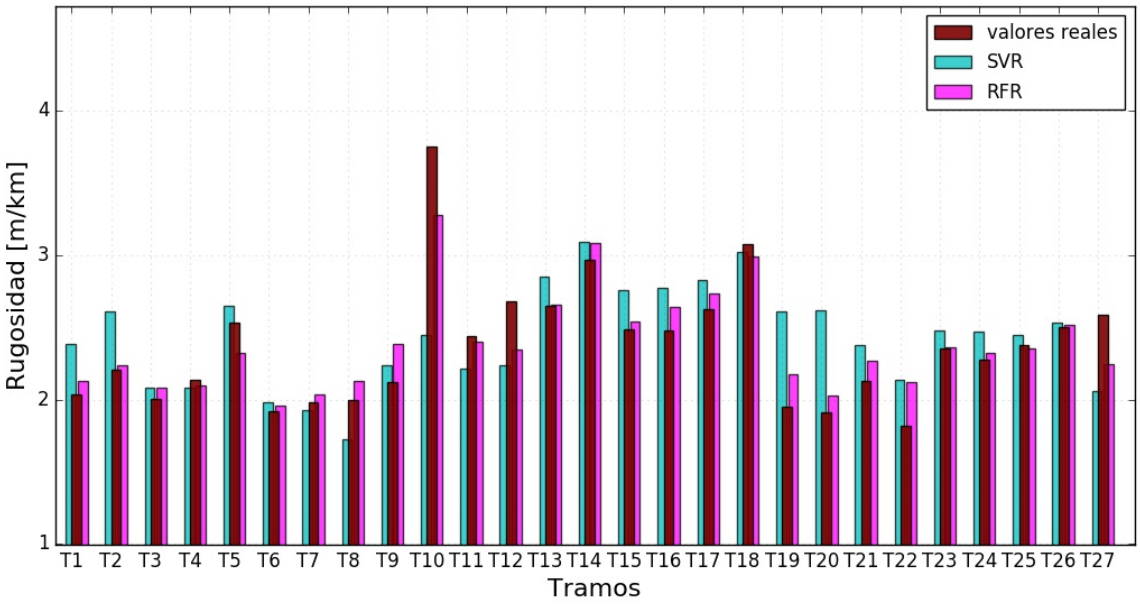


Figura 4.8: Predicción para el 3er. año con los valores optimizados.

Para los tramos 4, 5, 10, 11, 12, 18, 25 y 27 con *Random Forest* y para los tramos 4, 7, 8, 10, 11, 12, 18 y 27 con *Support Vector Machine*, dieron valores inferiores al último conocido, por lo que según la predicción, la rugosidad mejoraría al pasar los años.

Forzando mediciones ascendentes

Si se observan los datos, si bien se redujo el ruido por la polinómica de grado 3, las gráficas siguen teniendo pendientes descendientes; es decir, que de un año al siguiente, mejora la rugosidad sin realizar obra alguna. Esto no es correcto.
El siguiente paso será forzar a que las mediciones sean ascendentes. Si el IRI del año siguiente es menor que el actual, entonces el IRI siguiente tomará el valor del IRI actual; y así con las mediciones que le siguen.

Al igual que en el anterior análisis, para optimizar los parámetros, primero se predice el último año conocido.

En la figura 4.9 se pueden visualizar las curvas con ambos modelos. Y en la tabla 4.11 se muestra los errores cuadráticos de los mismos, con los parámetros optimizados.

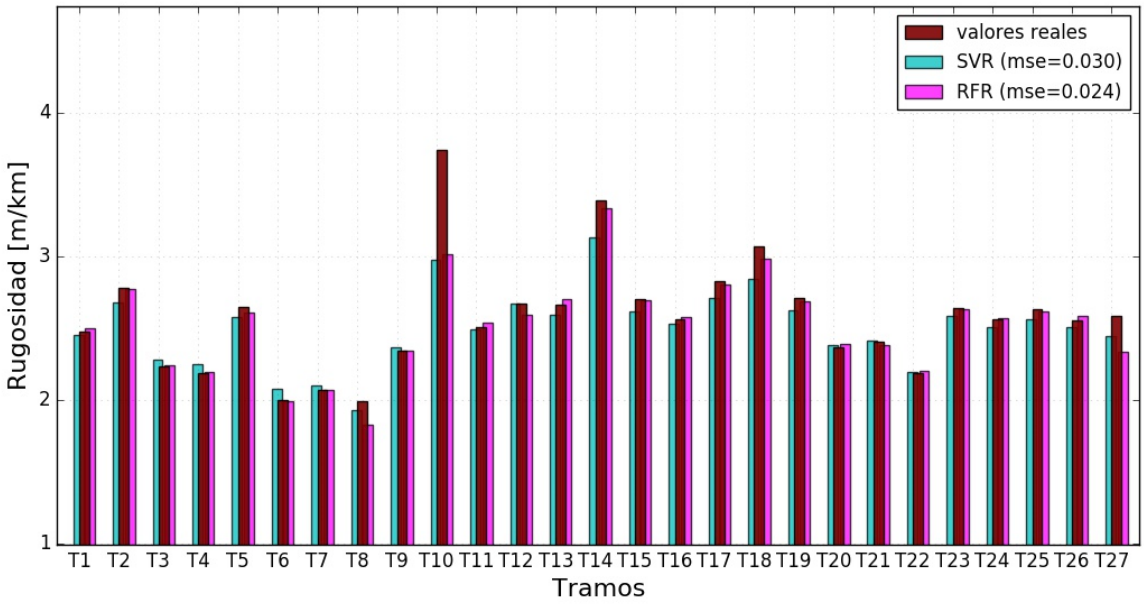


Figura 4.9: Predicción del último año conocido forzando mediciones ascendentes.

Modelo	Parámetros	Error $[m/km]^2$
SVR	$C = 1$ - $\varepsilon = 0.005$	0.030
RFR	700 árboles	0.024

Tabla 4.11: Errores para predicción del último año conocido forzando mediciones ascendentes.

Ahora se intentará predecir a 3 años futuros con los valores optimizados. Se entrena con ambos métodos. En la figura 4.10 se visualizan ambos modelos.

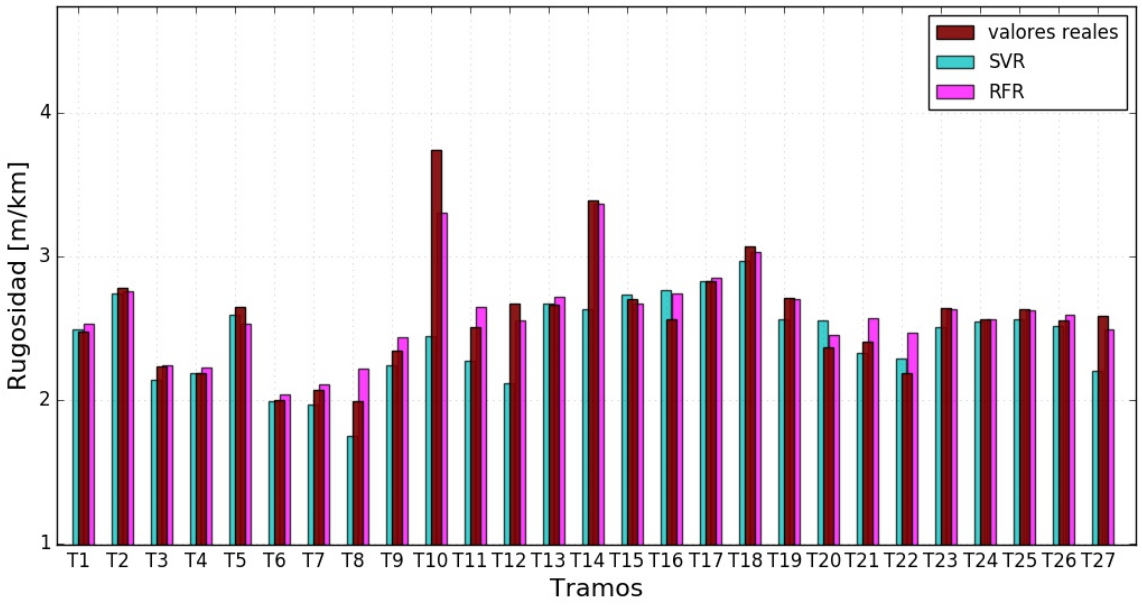


Figura 4.10: Predicción para el 3er. año forzando mediciones ascendentes.

Ahora, con *Random Forest*, los tramos 2, 5, 10, 12, 14, 15, 18, 25, 27 según la predicción, la rugosidad mejoraría al pasar los años. Con *Support Vector Machine* sucedería lo mismo para los tramos 2, 3, 5, 7, 8, 9, 10, 11, 12, 14, 18, 19, 21, 24, 25, 26 y 27. El modelo sigue no siendo satisfactorio.

Entrenando con datos de validación

¿Se podría entrenar con los datos de testeo? En realidad, lo que se ha hecho hasta ahora es predecir qué ocurriría para un año determinado. Particularmente, para un vector de entrada generado donde el año supera en 3 al último año conocido. Se validó el modelo con un vector que se generó como:

- último año de la medición + 3
- última deflexión
- (último transito liviano)$\cdot 1.02^3$
- (último transito medio)$\cdot 1.02^3$

- (último transito pesado)$\cdot 1.02^3$

No se planteó la posibilidad de cómo iría "empeorando" la rugosidad. Para esto se necesitaría pasar la rugosidad del año anterior como *input*.
Se agregó la columna con el IRI del año anterior, prediciendo año tras año (generando un nuevo vector hasta llegar a los años que se pretende predecir, en este caso 3), testeando con el IRI predicho. Además, en lugar de descartar años anteriores a la mejora, lo que se hizo fue dividir ese tramo en dos, colocándolo como un nuevo tramo desde el año cero; de esta manera no se pierden datos de entrenamiento.

Con *Random Forest*, la curva que se obtuvo como resultado fue óptima. Como se aprecia en la figura 4.11, el resultado predicho se encuentra por arriba de la curva. No obstante, para los tramos 29, 30, 31 y 32, según la predicción, la rugosidad mejoraría al pasar los años.

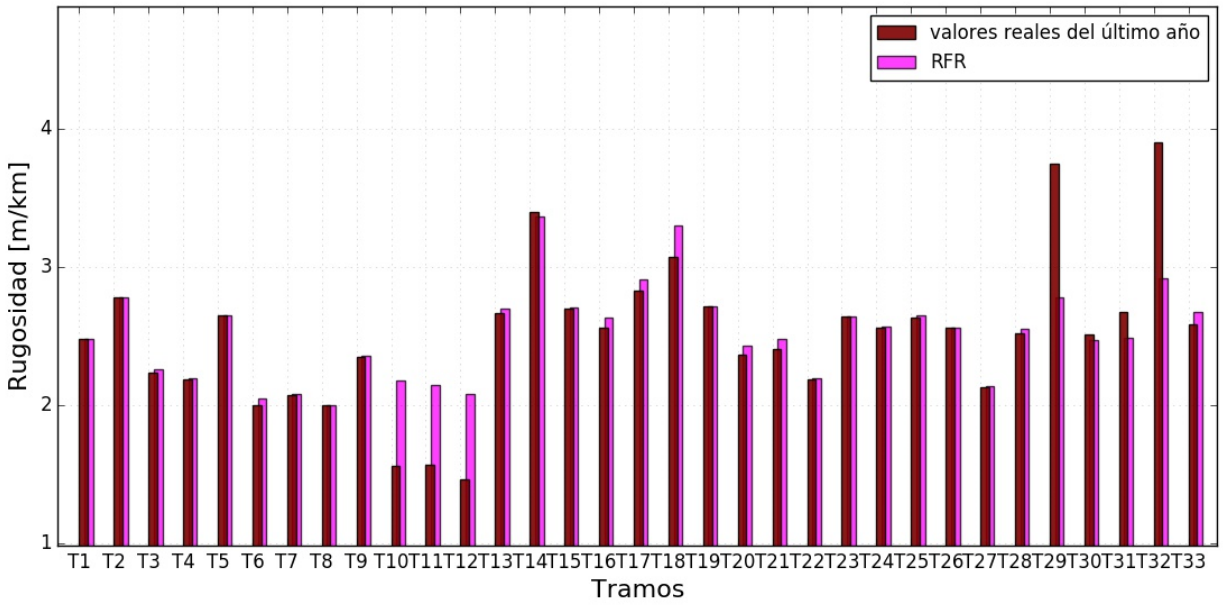


Figura 4.11: Predicción para el 3er. año con *Random Forest Regression* entrenando con los datos de validación.

El tramo 14 (ver tabla 4.12) se dividió en tres nuevos tramos, quedando como tramo 14 solo el año 2009 y 2010. El tramo 29 se generó a partir de la mejora del tramo 9, en el año 2010; quedando como tramo 9 desde el año 2010 en adelante, y como tramo 29 una

única medición. Para estos tramos, el modelo predicho no trabaja de manera eficiente, puesto que se está pretendiendo predecir a 3 años cuando se cuenta con menos cantidad de años para aprender.

Tramo	Año	Deflexión	Tránsito Liviano	Tránsito Medio	Tránsito Pesado	IRI
14	2009	89	1293	42	333	3.500
14	2010	89	1293	42	333	3.899
14	2011	89	1293	42	333	3.119
14	2012	44	1814	148	338	3.247
14	2013	44	1494	62	396	3.434
14	2014	44	1494	62	396	2.802
14	2015	44	1494	62	396	2.987

Tabla 4.12: Datos del tramo 14.

La figura 4.12 muestra el resultado del *Support Vector Machine Regression* con los parámetros anteriormente optimizados. La curva mejora, pero los resultados no siguen siendo satisfactorios ya que para algunos tramos mejoraría la rugosidad al pasar los años.

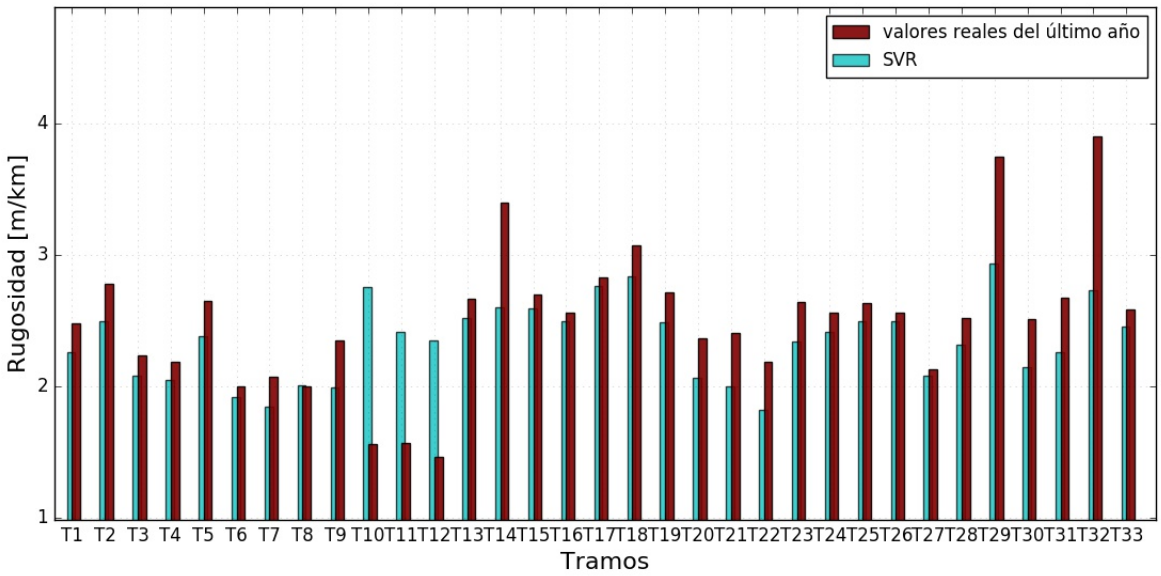


Figura 4.12: Predicción para el 3er. año con *Support Vector Machine Regression* entrenando con los datos de validación.

El ir agregando el vector predicho como validación, genera ruido exponencial, ya que las predicciones de cada año desconocido introducen un error intrínseco que va a ser acumulado al emplearlo como entrada para la siguiente predicción.

Los resultados obtenidos demuestran la eficacia del método *Random Forest* con respecto

a *Vector Support Machine*. Para *Vector Support Machine*, los puntos en los que el error de interpolación es menor que ε no son *vectores soporte* y no forman parte de la solución. Al parecer, si bien se minimizó el ruido para este último modelo, el método fue a lo seguro. Sin embargo, *Random Forest* es eficaz para eliminar el ruido en los datos de entrada del modelo debido a que construye muchos árboles usando un subconjunto de las variables de entrada disponibles y sus valores, contiene árboles de decisión subyacentes que omiten la variable generadora de ruido.

Discusión y conclusiones

Se ha logrado de manera óptima un resultado satisfactorio. Partiendo de una base de datos que constaba de 27 tramos homogéneos evaluados desde el año 2009 al 2015, se conformó una totalidad de 189 conjunto de datos.

Se comenzó prediciendo la rugosidad para el último año conocido. Resultando con un menor error de predicción el regresor *Vector Support Machine*. El siguiente paso fue filtrar las mejoras estructurales no documentadas, eliminado las mediciones anteriores a las mismas, ya que había ciertos picos para los que el regresor se comportaba de manera incorrecta. Se redujo notoriamente el error cuadrático medio de ambos modelos. Se desglosó la variable TMDA en tránsito liviano, medio y pesado, lográndose mejorar levemente la predicción.

Se intentó hacer una validación *Leave-one-out* buscando predecir no muy satisfactoriamente un tramo entero. Se entrenó con acumulados y manteniendo el IRI del año anterior, llegando a la conclusión de que esta validación no tiene sentido lógico en la vida útil de un tramo homogéneo, puesto que lo que se pretende predecir es el estado de deterioro de circulación para un año futuro determinado, y no la totalidad del tramo. Si existe el tramo, se cuenta con los datos iniciales del proyecto, por lo tanto se conoce el tránsito, deflexión, rugosidad inicial; por lo que no es un dato desconocido a predecir.

El siguiente paso fue la predicción a 3 años futuros. Es decir, se generaron vectores de entrada como:

- último año de la medición + 3
- última deflexión

- (último transito liviano)$\cdot 1.02^3$
- (último transito medio)$\cdot 1.02^3$
- (último transito pesado)$\cdot 1.02^3$

Además, se comenzó reduciendo posibles ruidos de medición. Para esto, se aproximó por una polinómica de grado 3 para generar la rugosidad de esos años eliminados. Luego de optimizar los valores prediciendo al último año conocido, tanto *Random Forest* como *Support Vector Machine* mostraron que en ciertos tramos la rugosidad mejoraría al pasar los años. Esto equivale a decir que los resultados del predictor estaban por debajo del IRI correspondiente al último año medido.

Al observar que existían pendientes descendentes, se forzó a que las mediciones sean ascendentes: si el IRI del año siguiente era menor que el actual, entonces el IRI siguiente tomaría el valor del IRI actual; y así con las mediciones que le siguieran. Luego de entrenar con *Vector Support Machine* y *Random Forest* siguieron existiendo rugosidad que estaban por debajo del último año conocido.

Se logró llegar a un resultado satisfactorio cuando fue necesario considerar los datos de testeo como entrenamiento, es decir, no solo predecir qué sucede dentro de 3 años, sino año tras año. Para lograr esto, se pasó la rugosidad del año anterior como entrada. Y para optimizar el modelo, en lugar de descartar años anteriores a la mejora, lo que se hizo fue dividir ese tramo, colocándolo como nuevo tramo desde el año cero; de esta manera no se perdió ningún dato de entrenamiento.

Random Forest trabaja de manera óptima para la predicción de deterioro de diferentes tramos. No se puede decir lo mismo de los *Support Vector Machine*; los cuales se estiman que darían mejor resultado si se tuviera más datos de entrenamientos (más años y más tramos). No obstante, se podría disponer de los datos de vialidad nacional para ampliar este estudio y reveer los resultados acá obtenidos.

Como trabajo futuro se adaptaría el método para la predicción de otras variables que influyen en la evolución de los deterioros de los pavimentos. Como por ejemplo **fisuras**,

que es la discontinuidad en la capa de rodamiento. Este deterioro disminuye la resistencia de la capa afectada y evoluciona hacia deterioros más serios como la rugosidad, desprendimientos, baches. Constituyen una potencial vía de acceso de agua hacia las capas inferiores.

Otra variable a analizar sería **ahuellamiento**, que son los cambios del perfil transversal del camino, respecto al perfil original construido o pretendido. Estas variaciones del perfil transversal se encuentran en las huellas y aparecen debido a la deformación permanente causada por los vehículos pesados, en todas las capas de la estructura. El grado de deformación tiene una incidencia fundamental en las condiciones funcionales (para el usuario) y estructurales (para el ingeniero) del pavimento existente. Esto ocasiona acumulación del agua en las huellas, produciendo pérdida del control del vehículo, inseguridad en las maniobras, falta de confort, lo que conlleva a un gran riesgo de accidentes [1].

Estas variables impactan sobre la rugosidad de manera indirecta. La predicción de estos deterioros en si mismo resulta un objetivo deseado. Colateralmente su inclusión como variable de entrada en el modelo de rugosidad planteará seguramente la mejora del mismo.

Bibliografía

[1] Material de apoyo del curso "Evaluación de Calzada". *Evaluación de calzada*. FCEIA-IMAE, 2016.

[2] Tom Michell. *Tom Michell*, 1998.

[3] Wikipedia. Aprendizaje automático. `https://es.wikipedia.org/wiki/Aprendizaje_automatico`. [Accedido en marzo de 2017].

[4] Raul E. Lopez Briega. Machine Learning con Python. `http://relopezbriega.github.io/blog/2015/10/10/machine-learning-con-python/`, 2015. [Accedido en marzo de 2017].

[5] Juan José Martín Guareño. Support vector regression: Propiedades y aplicaciones. Master's thesis, Facultad de Matemáticas, Departamento de Estadística e Investigación Operativa de la Universidad de Sevilla, 2016.

[6] John Goddard, Sergio Gerardo de los Cobos Silva, Blanca Rosa Perez Salvador, and Miguel Angel Gutierrez Andrade. Un algoritmo para el entrenamiento de máquinas de vector soporte para regresión. *Revista de Matematica: Teoría y Aplicaciones 2000 7(1-2) : 107–116*, 2000. issn: 1409-2433.

[7] Alex J. Smola and Bernhard Schölkopf. *A Tutorial on Support Vector Regression*, 2003.

[8] Support Vector Machine Regression. `http://kernelsvm.tripod.com/`. [Accedido en julio de 2016].

[9] Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. *Support Vector Regression Machines*. Bell Labs and Monmouth University Department of Electronic Engineering West Long Branch, 1996.

[10] L. Breiman. *Bagging Predictors. Machine Learning 24 (2) pp. 123–140 (2001).*

[11] Adele Cutler. *Trees and Random Forests*. Mathematics and Statistics, Utah State University, 2013.

[12] Quora Ankit Sharma. How does random forest work for regression? `https://www.quora.com/How-does-random-forest-work-for-regression-1`, 2014. [Accedido en agosto de 2016].

[13] Material de apoyo a la cátedra Transporte III. *Auscultación de pavimentos*. Carrera Ingeniería Civil FCEIA UNR, 2015.

[14] Hugo Poncino, Marta Pagola, Oscar Giovanon, Mario Noste, and Jorge Tosticarelli. *Rugosidad de Pavimentos*. Publicación utilizada para dictado de cursos a Vialidades Provinciales, 1996.

[15] Scikit learn. Preprocessing data. `http://scikit-learn.org/stable/modules/preprocessing.html`. [Accedido en abril de 2017].