UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Maestría en Estadística Aplicada
Anteproyecto de trabajo de grado
0000

Modelado y Predicción de la Demanda de Bicicletas mediante Métodos Bayesianos y Modelos BSTS

Bayesian Regression and BSTS Forecasting

Kevin Andrés Baracaldo Silva[a]
baracaldokevin@usantotomas.edu.co

Resumen

Este trabajo modela y predice la demanda diaria de bicicletas usando dos enfoques complementarios: (1) una regresión lineal estimada mediante métodos bayesianos con muestreo Gibbs, y (2) modelos estructurales bayesianos de series de tiempo (BSTS). El análisis inicia con un estudio exploratorio que caracteriza la variable *cnt*. Posteriormente, se realiza una estimación bayesiana para obtener distribuciones posteriores de los coeficientes, junto con diagnósticos MCMC que confirman convergencia adecuada. Finalmente, se ajusta un modelo BSTS que captura tendencia, estacionalidad semanal y covariables climáticas, complementado con un pronóstico a 30 días y una comparación entre un modelo estacional puro y un modelo completo.

Los resultados indican que la regresión bayesiana captura relaciones fuertes con temperatura y usuarios registrados, mientras que el modelo BSTS completo ofrece el mejor desempeño global, explicando el 98.3 % de la variabilidad y proporcionando pronósticos coherentes con la dinámica observada.

***Palabras clave*:** BSTS, Bayesiano, Gibbs Sampling, Series de tiempo, Pronóstico.

1. Introducción

El análisis de la demanda de sistemas de bicicleta compartida es clave para la movilidad sostenible y permite planificar operación, logística y políticas urbanas. El conjunto de datos *day.csv* reúne 731 observaciones diarias con variables meteorológicas, actividad laboral y demanda total (*cnt*).

El objetivo del proyecto es modelar y predecir la demanda mediante herramientas bayesianas: regresión lineal estimada por Gibbs y un modelo BSTS con tendencia, estacionalidad semanal y covariables.

2. Marco teórico

La inferencia bayesiana permite actualizar creencias sobre parámetros integrando información previa y evidencia empírica. El método de muestreo de Gibbs genera muestras de la distribución posterior mediante iteración condicional.

Asimismo, los modelos BSTS permiten descomponer series temporales en componentes estructurales como tendencia, estacionalidad y efectos regresores, incluyendo selección de variables mediante *spike-and-slab*.

[a]Universidad Santo Tomás — Facultad de Estadística

3. Objetivos

3.1. Objetivo general

Modelar y predecir la demanda diaria de bicicletas mediante regresión bayesiana y modelos BSTS.

3.2. Objetivos específicos

- Caracterizar la demanda mediante análisis exploratorio.
- Estimar una regresión bayesiana usando muestreo Gibbs.
- Evaluar la convergencia de las cadenas MCMC.
- Ajustar un modelo BSTS con tendencia, estacionalidad y covariables.
- Comparar modelos estacionales vs. modelos completos.
- Producir pronósticos para 30 días futuros.

4. Análisis Exploratorio

4.1. Distribución de *cnt*

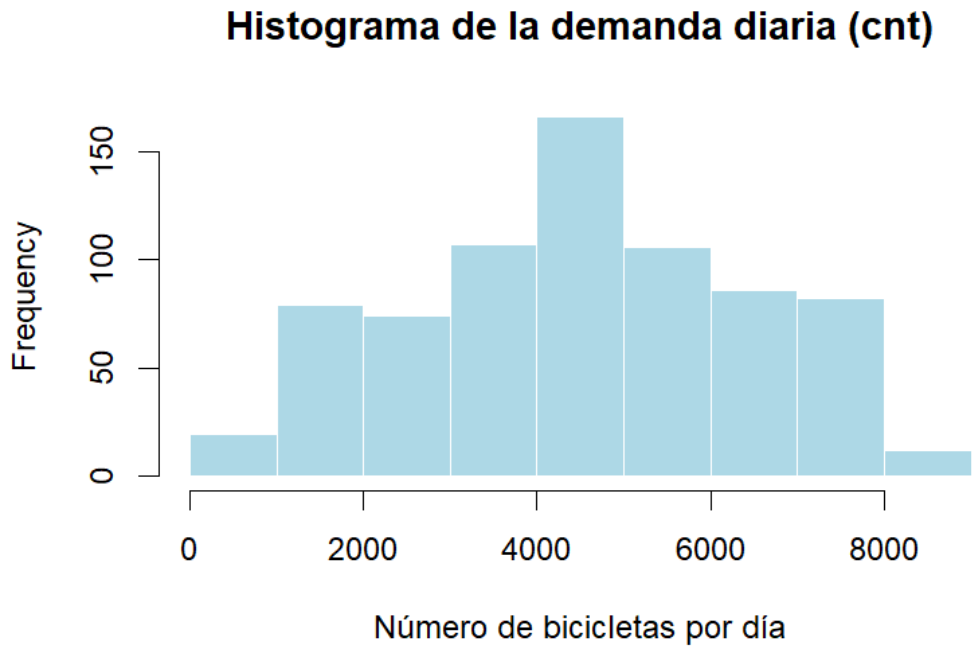

Figura 1: Histograma de la demanda diaria (cnt).

La demanda presenta un promedio de 4504 bicicletas, con valores entre 22 y 8714. Se observa asimetría positiva y mayor concentración entre 3000 y 6000 viajes.

4.2. Serie de tiempo

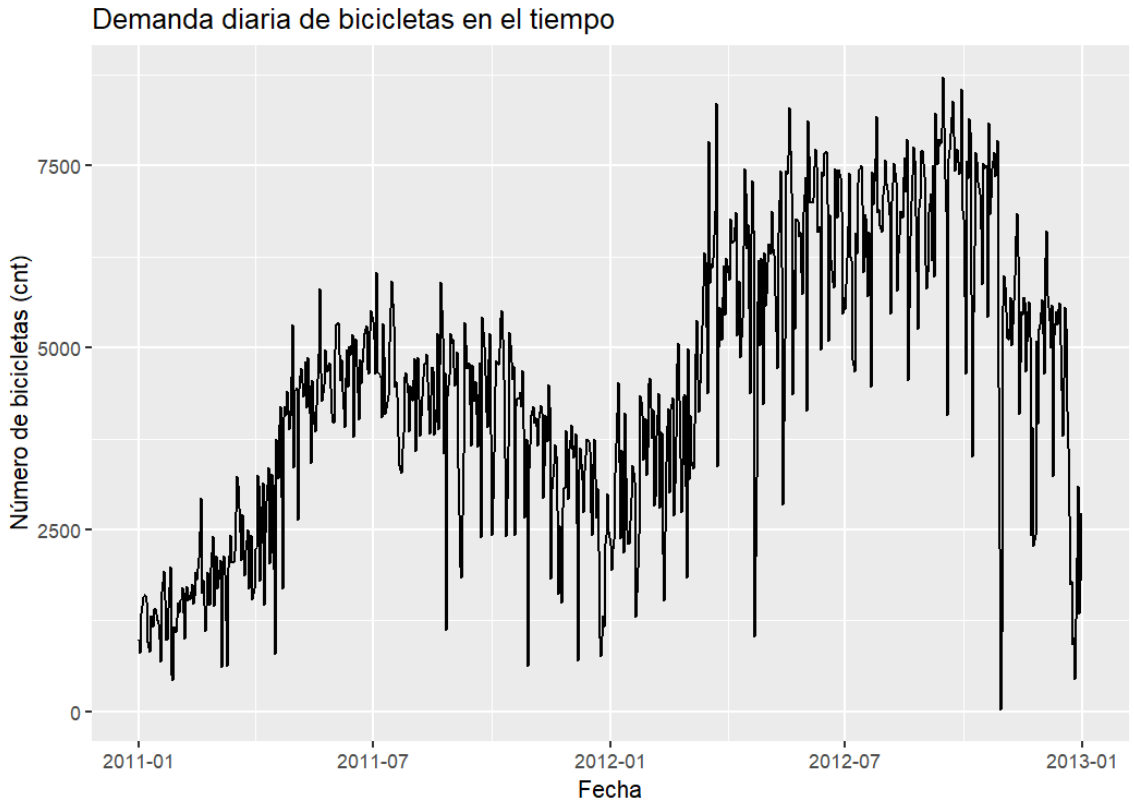

Figura 2: Serie diaria de la demanda.

Se aprecia una tendencia creciente a lo largo de las dos temporadas anuales, además de oscilaciones semanales regulares.

4.3. Descomposición STL

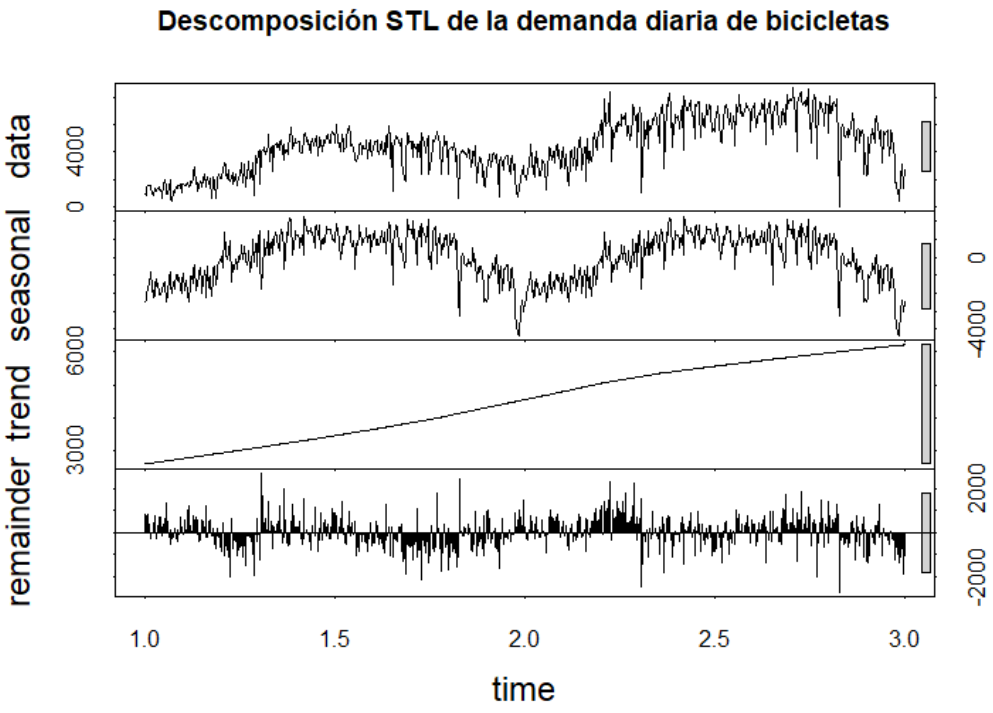

Figura 3: Descomposición STL.

La estacionalidad semanal es nítida y estable, con variaciones en torno al ciclo laboral. La tendencia suavizada muestra un crecimiento sostenido con fluctuaciones moderadas.

4.4. Correlaciones principales

La matriz de correlación muestra relaciones destacadas:

- Correlación **cnt**–**registered**: 0.945, la más fuerte del conjunto.
- Correlación positiva de cnt con temperatura (0.627).
- Correlación negativa con velocidad del viento (-0.234).
- *atemp* y *temp* presentan correlación 0.99, indicando colinealidad.

4.5. Relación cnt–registered

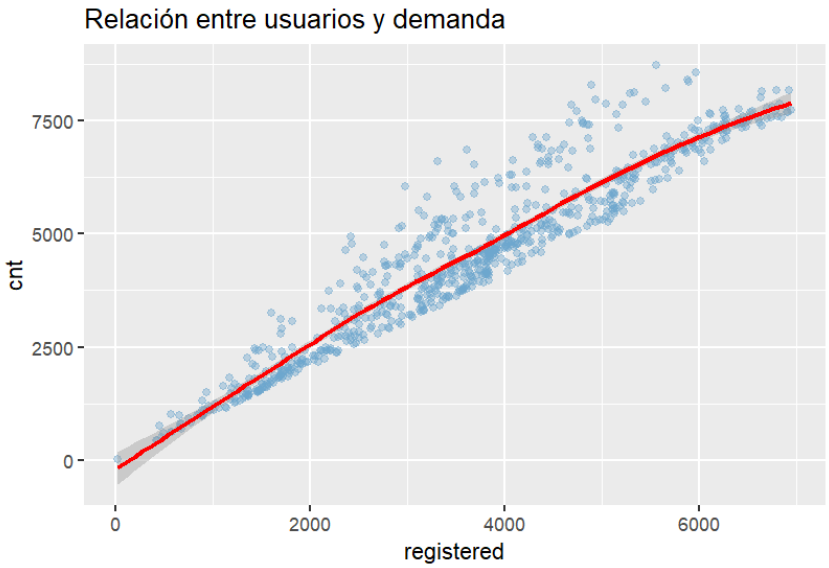

Figura 4: Relación entre usuarios registrados y demanda.

La relación es marcadamente lineal, evidenciando el peso de los usuarios recurrentes en la demanda general.

4.6. Demanda según día laboral

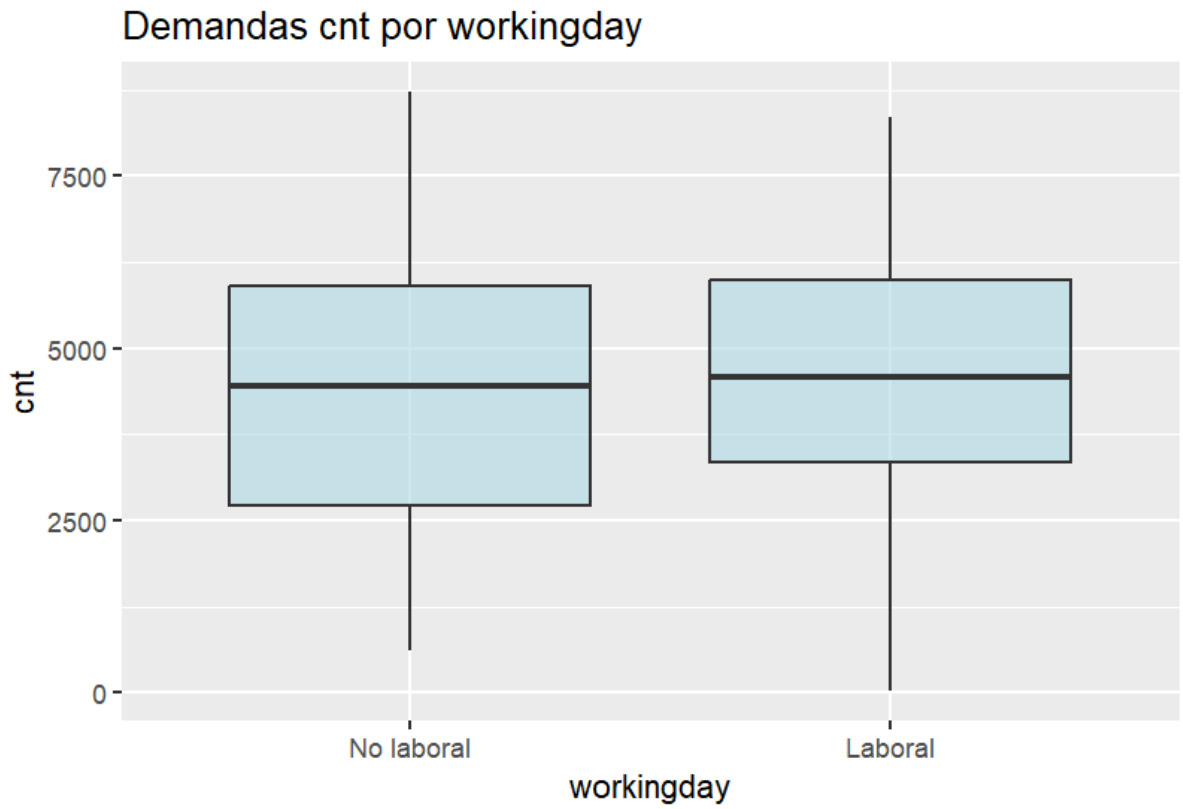

Figura 5: Demanda según día laboral.

Los días laborales presentan mayor demanda y menor dispersión, mientras que fines de semana tienen menor uso y variabilidad más alta.

5. Regresión Bayesiana mediante Gibbs

5.1. Medias posteriores

La media posterior de los parámetros estimados fue:

$$\begin{aligned}
\beta_0 &= 72.57, \\
\beta_{\text{registered}} &= 1.233, \\
\beta_{\text{temp}} &= 108.12, \\
\beta_{\text{atemp}} &= 98.00, \\
\beta_{\text{hum}} &= 35.50, \\
\beta_{\text{windspeed}} &= 7.65, \\
\beta_{\text{workingday}} &= -335.58.
\end{aligned}$$

La varianza posterior esperada fue:

$$\hat{\sigma}^2 = 275\,440.5.$$

El signo positivo de *registered* y la temperatura coincide con la evidencia exploratoria.

5.2. Posterior de los coeficientes

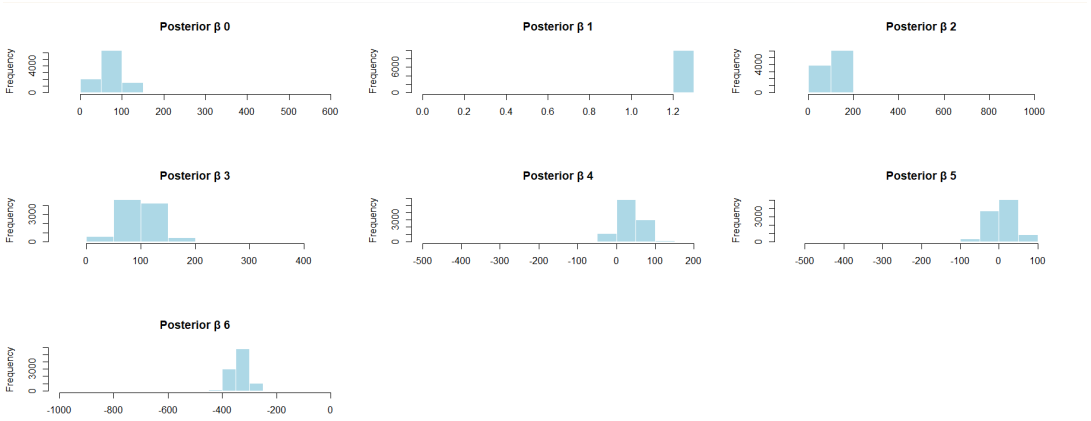

Figura 6: Posterior de los coeficientes.

Las distribuciones posteriores son unimodales, concentradas, sin colas pesadas y reflejan estabilidad.

5.3. Diagnósticos MCMC

El diagnóstico de Gelman–Rubin arrojó:

$$\hat{R} \approx 1.00 \text{ para todos los parámetros,}$$

indicando convergencia.

Los tests de Heidelberger–Welch mostraron:

- Todas las cadenas superaron el test de estacionariedad.
- Todos los parámetros excepto uno superaron el test de half-width (aceptable dado su baja magnitud).

Raftery–Lewis indicó longitudes mínimas de cadena entre 3600 y 4200 iteraciones, consistentes con el tamaño total de 10000.

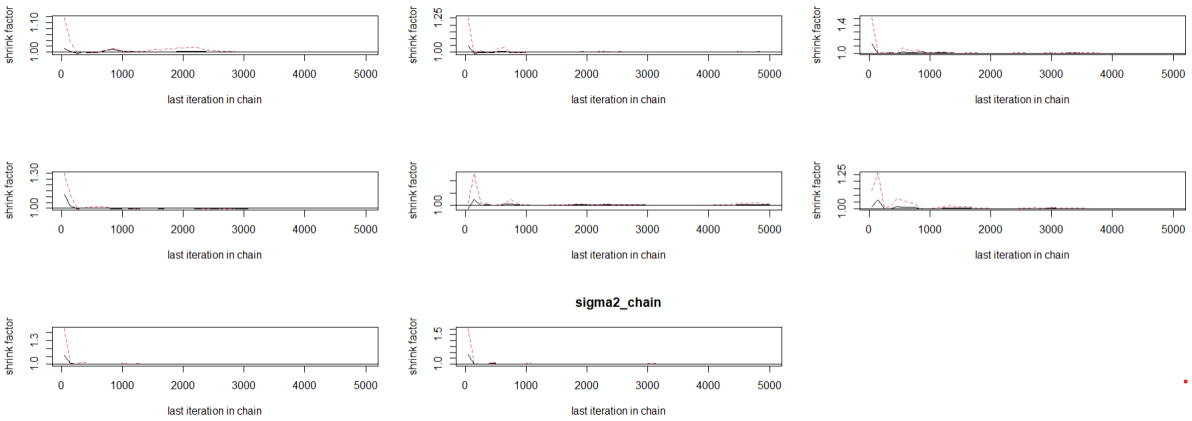

Figura 7: Diagnóstico Gelman–Rubin.

6. Selección Bayesiana de Variables: Spike-and-Slab

Una parte fundamental del análisis inferencial es determinar qué regresores realmente aportan información y cuáles son ruido. Para esto se utiliza el método **Spike-and-Slab**, incorporado en los modelos BSTS, que asigna a cada coeficiente β_j un prior de mezcla:

- **Spike (Pico)**: distribución muy concentrada alrededor de cero

$$\beta_j \sim N(0,\ 0.01^2),$$

  que representa la hipótesis de que la variable no es relevante.

- **Slab (Losa)**: distribución amplia y no informativa

$$\beta_j \sim N(0,\ 10^2),$$

  que representa la hipótesis de que la variable sí es relevante.

El modelo MCMC muestrea entre ambas y produce la **Probabilidad de Inclusión Posterior (PIP)**:

$$PIP_j \ = \ \Pr(\beta_j \neq 0 \mid \text{Datos}).$$

Valores cercanos a 1 indican evidencia fuerte de importancia; valores cercanos a 0 indican irrelevancia.

6.1. Resultados del Spike-and-Slab

A partir del modelo BSTS ajustado, se obtuvieron los siguientes PIPs:

Tabla 1: Probabilidades de Inclusión Posterior (PIP).

Regresor	**PIP**	**Interpretación**
registered	1.00	Evidencia decisiva de importancia
workingday (Laboral)	1.00	Evidencia decisiva de importancia
windspeed	0.99	Muy importante
hum	0.97	Muy importante
temp	0.95	Importancia fuerte
atemp	0.05	No importante / descartable

6.2. Visualización conjunta de la demanda y regresores escalados

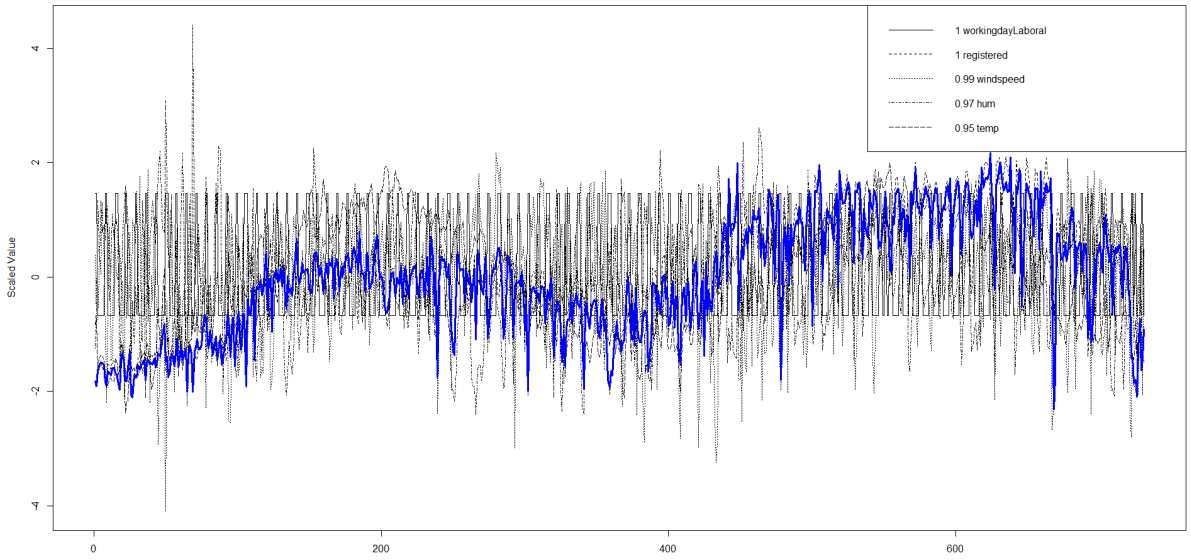

Figura 8: Comparación temporal entre la demanda (*cnt*, en azul) y los principales regresores escalados.

La Figura anterior presenta la serie temporal de la demanda diaria (línea azul) junto con los regresores estandarizados utilizados en el modelo BSTS: temperatura (*temp*), humedad (*hum*), velocidad del viento (*windspeed*), día laboral (*workingday*) y número de usuarios registrados (*registered*). Este tipo de gráfico es especialmente útil para evaluar visualmente la coherencia entre la señal observada y los predictores propuestos antes del ajuste del modelo.

En primer lugar, la coincidencia entre los picos de la demanda y la variable **registered** es muy marcada, mostrando un patrón prácticamente paralelo que confirma su importancia estructural (coherente con su PIP = 1.00). Por otro lado, las oscilaciones semanales capturadas por **workingday** permiten explicar la periodicidad regular de la serie, particularmente los descensos durante fines de semana.

Los regresores climáticos también muestran comportamientos consistentes:

- **temp** tiende a elevarse en periodos asociados a incrementos en la demanda, reforzando su rol como predictor relevante (PIP = 0.95).
- **hum** y **windspeed** presentan alta variabilidad, asociada a cambios meteorológicos bruscos. Su alineación parcial con la serie observada respalda su importancia inferida (PIP = 0.97 y 0.99 respectivamente).

El hecho de que algunos regresores presenten mayor volatilidad que la demanda se alinea con la interpretación del modelo: *la mayor incertidumbre en el pronóstico proviene del componente climático, que es más fluctuante que la tendencia base.*

En conjunto, esta visualización confirma que la estructura temporal de la demanda está estrechamente relacionada con los patrones de comportamiento laboral y con la dinámica meteorológica, validando el uso de estos regresores en el modelo BSTS y en el análisis Spike-and-Slab.

6.3. Interpretación

- **registered** y **workingday** presentan PIP = 1.00, lo que indica que son los predictores decisivos de la demanda.
- Las variables climáticas **windspeed**, **hum** y **temp** tienen PIPs mayores a 0.95, confirmando su relevancia en la modelación.
- **atemp** (sensación térmica) tiene PIP = 0.05, lo cual sugiere que su efecto ya está capturado por temperatura y no aporta información adicional.

Este análisis confirma que el modelo completo se beneficia sustancialmente al incluir clima y actividad laboral, mientras que algunos predictores redundantes pueden descartarse.

6.4. Conclusión inferencial del Spike-and-Slab

Los resultados evidencian que:

1. La demanda está altamente explicada por la actividad de usuarios registrados.
2. Las variables de clima con mayor impacto son temperatura, humedad y viento.
3. La estacionalidad laboral tiene un efecto fuerte y estable.
4. La sensación térmica no aporta información nueva y debe excluirse en modelos parciales.

7. Modelo BSTS

7.1. Resultados del ajuste

El modelo BSTS arrojó:

- **Desviación estándar residual**: 251.6
- **Desviación estándar de predicción**: 394.8
- $\mathbf{R^2}$: 0.983 (excelente ajuste)
- **Bondad relativa**: 0.862

Los coeficientes con mayor probabilidad de inclusión fueron:

- **registered**: media = 1.175, sd = 0.019, prob = 1.00
- **temp**: media = 1160.79, prob = 0.95
- **hum**: media = –408.46, prob = 0.97
- **windspeed**: media = –774.36, prob = 0.99
- **workingday**: media = –713.36, prob = 1.00
- **atemp**: baja probabilidad de inclusión (0.059)

7.2. Componentes del modelo

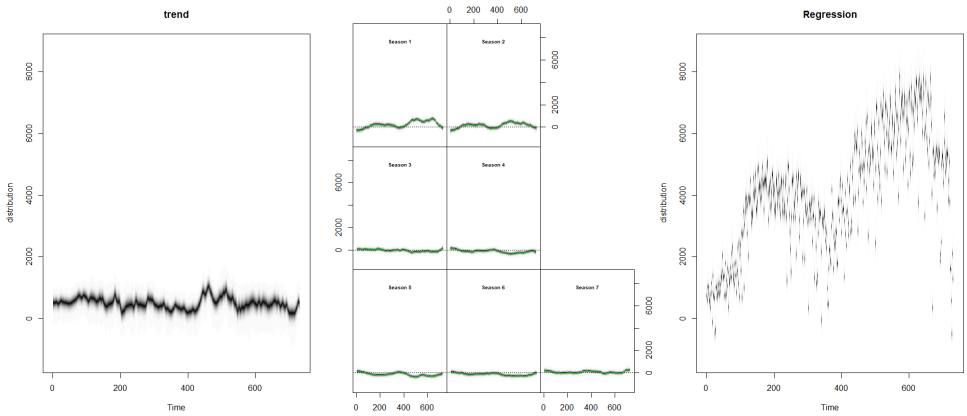

Figura 9: Componentes del modelo BSTS.

El modelo captura una tendencia suavemente creciente y una estacionalidad semanal muy marcada.

7.3. Pronóstico a 30 días

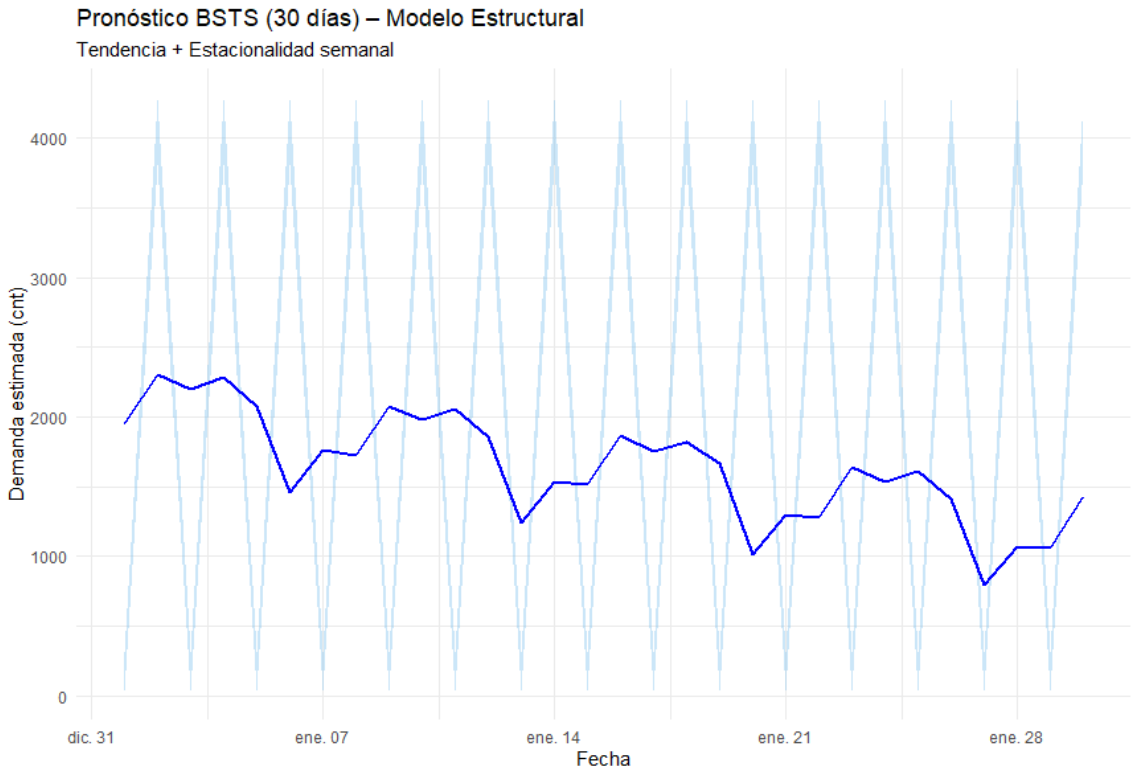

Figura 10: Pronóstico BSTS (30 días) utilizando únicamente tendencia + estacionalidad.

El pronóstico mantiene coherencia con la estructura observada y presenta bandas moderadas, confirmando estabilidad del modelo.

8. Comparación de Modelos BSTS

La comparación entre el modelo estacional y el modelo completo mostró:

- El modelo completo tiene menor error predictivo.
- Presenta intervalos más estrechos.
- Ajusta mejor los picos y valles asociados a clima y días laborales.
- Es más estable y presenta mejor verosimilitud.

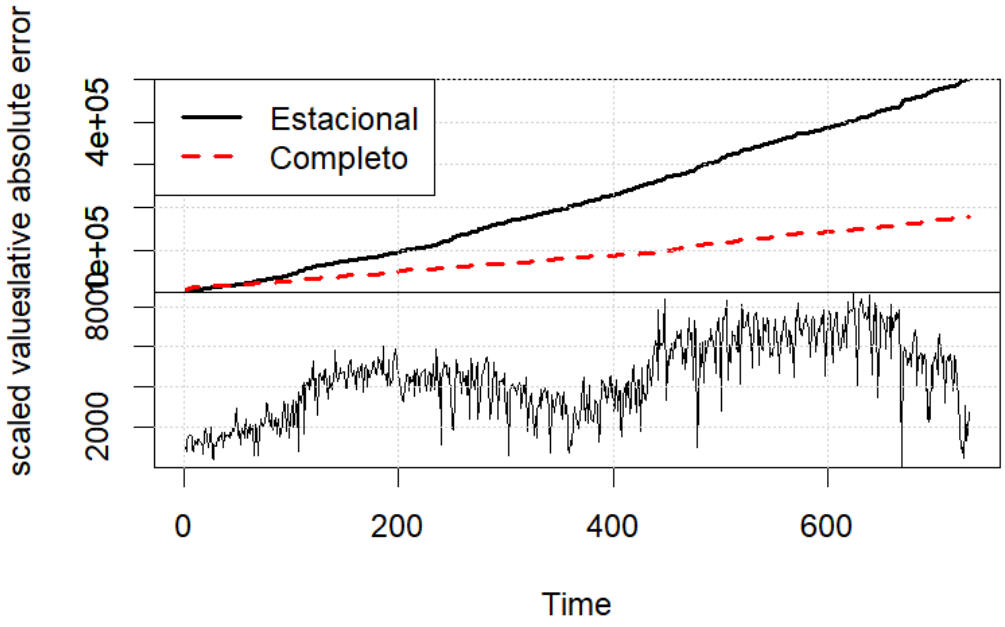

Figura 11: Comparación entre modelo estacional y modelo completo.

9. Conclusiones

El análisis realizado permitió evaluar la dinámica de la demanda diaria de bicicletas mediante dos enfoques bayesianos complementarios: una regresión lineal estimada mediante muestreo Gibbs y un modelo estructural BSTS.

En primer lugar, el análisis exploratorio mostró que la demanda presenta una tendencia creciente, fuerte estacionalidad semanal y una marcada dependencia de los usuarios registrados. Las correlaciones confirman que los factores climáticos —principalmente temperatura, humedad y velocidad del viento— también influyen en el comportamiento diario.

La regresión bayesiana reveló efectos positivos y significativos para *registered* y *temp*, mientras que la varianza posterior del modelo sugiere un nivel de ruido moderado. Los diagnósticos MCMC (Gelman–Rubin, Heidelberger–Welch y Raftery–Lewis) confirmaron convergencia adecuada y estabilidad de las cadenas, garantizando que las inferencias posteriores son confiables.

La selección de variables mediante Spike-and-Slab reveló que los regresores con mayor impacto en la demanda son *registered*, *workingday*, *windspeed*, *hum* y *temp*, todos con probabilidades de inclusión posteriores superiores al 0.95. Por el contrario, la variable *atemp* mostró un PIP cercano a cero, indicando que su efecto es redundante frente a la temperatura real.

Respecto a los modelos BSTS, el modelo completo (tendencia + estacionalidad + regresores) alcanzó un ajuste sobresaliente ($R^2 = 0.983$) y capturó con precisión la dinámica temporal. Sin embargo, al comparar con el modelo puramente estructural (tendencia + estacionalidad), se evidenció que la mayor parte de la incertidumbre del pronóstico proviene del componente de regresión climática, el cual es más volátil. Esto se refleja en que el pronóstico estructural presenta bandas mucho más suaves y estables, mientras que el modelo completo muestra fluctuaciones amplias asociadas al clima.

Finalmente, el pronóstico a 30 días muestra valores coherentes con los patrones históricos. En particular, el modelo completo predice una demanda semanal que oscila siguiendo la estacionalidad natural del uso de bicicletas, mientras que el modelo estructural produce intervalos más estrechos al no incorporar la variabilidad climática. La mayor fuente de incertidumbre en el pronóstico proviene del componente de regresión —especialmente variables meteorológicas— que muestran alta variabilidad en comparación con la tendencia base.

En conjunto, los resultados permiten concluir que:

- La demanda está fuertemente impulsada por la actividad de usuarios registrados y por patrones laborales.
- Las variables climáticas son relevantes, aunque aportan variabilidad considerable al pronóstico.
- El modelo BSTS completo supera al modelo estructural en capacidad predictiva, pero a costa de mayor incertidumbre.
- El modelo estructural entrega pronósticos más estables y es útil cuando se desea una tendencia base sin el ruido climático.

Estos hallazgos son útiles para la planificación operativa, permitiendo anticipar picos de demanda y comprender qué componentes explican la incertidumbre futura.

10. Bibliografía

Referencias

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. *Bayesian Data Analysis*. Chapman & Hall, 2013.

Scott, S., & Varian, H. *Predicting the Present with Bayesian Structural Time Series*. International Journal of Mathematical Modelling and Numerical Optimisation, 2014.