Análisis estadístico del modelo de Holt-Winters: Aplicación a la contaminación de aire por pm2.5 de Lima, Perú

Statistical analysis of the Holt-Winters model: Application to air pollution data for pm2.5, Lima, Peru

A análise estatística do modelo de Holt-Winters: Aplicação de poluição do ar para pm2.5, Lima, Peru

Elmis García Zare[1], José D. Bermúdez Edó[2]

Resumen

Se analizó los datos de monitoreo de material particulado (pm2.5) del Programa Nacional de Vigilancia Sanitaria de Calidad del Aire, Lima – Callao, utilizando el modelo de Holt-Winters y el análisis estadístico propuesto en Bermúdez et al. (2007). Esta metodología se basa en la estimación de los parámetros de suavizado y las condiciones iniciales mediante el método de máximo verosimilitud, además de los pronósticos puntuales y los intervalos de predicción, teniendo en cuenta algunos aspectos relativos a la distribución normal multivariante.

Palabras clave: Holt-Winters aditivo, material particulado, pronóstico de series de tiempo, suavizado exponencial.

Abstract

Monitoring data of particulate matter (pm2.5) National Sanitary Vigilance Program Air Quality, Lima-Callao, using the Holt-Winters model and statistical analysis proposed Bermudez et al. (2007) was analyzed. This methodology is based on the estimation of the smoothing parameters and initial conditions using the maximum likelihood method, besides the point forecasts and prediction intervals, taking into account some aspects of multivariate normal distribution.

Keywords: Holt-Winters additive, particulate matter, time series forecasting, exponential smoothing.

Resumo

Os dados de monitoramento de material particulado (pm2,5) Programa Nacional de Vigilância sanitária da qualidade do ar, Lima - Callao, usando o modelo de Holt-Winters e análise estatística proposto Bermudez et al. (2007) foi analisada. Esta metodologia baseia-se na estimativa dos parâmetros de alisamento e condições iniciais utilizando o método de máxima verossimilhança, além das previsões pontuais e intervalos de predição, tendo em conta alguns aspectos da distribuição normal multivariada.

Palavras-chave**:** Holt-Winters aditivo material particulado, previsão de séries temporais, suavização exponencial.

Introducción

Los métodos de predicción mediante el suavizado exponencial se basan en la atenuación de los valores de la serie de tiempo, obteniendo el promedio de estos de manera exponencial; es decir, los datos se ponderan dando un mayor peso a las observaciones más recientes y uno menor a las más antiguas. Dentro de las técnicas de suavización exponencial (Gardner, 2006; Hyndman et al., 2008; Ord et al., 1997), las

[1] Universidad Nacional de Trujillo, Perú

[2] Universitat de Valencia, España

Recibido, 4 de setiembre de 2015
Aceptado, 5 de noviembre de 2015

más utilizadas son: Suavización Exponencial Simple, Suavización Exponencial Doble (Método de Brown), Suavización Exponencial Ajustada a la tendencia (Método de Holt), y la Suavización Exponencial Triple o modelo de Holt-Winters.

El modelo de Holt-Winters, considera que la serie se puede descomponer en todos o algunos de los siguientes componentes: a) tendencia; b) factor cíclico; c) estacionalidad y d) componente irregular (Jiménez et al., 2006) y según los métodos de descomposición, las series son el resultado de la integración de esos cuatro componentes, bien de modo aditivo (las fluctuaciones no se ven afectadas por la tendencia) o de modo multiplicativo (las fluctuaciones varían con la tendencia). Tal procedimiento implica determinar valores iniciales y parámetros de suavizado cuya elección es de importancia, discutido esto por Chatfield and Yar (1988). Es usual utilizar métodos como razón de la media móvil, para el caso de los valores iniciales de estacionalidad; y ajustar a una recta la serie desestacionalizada, en el caso de los valores de nivel y tendencia tal como lo indica Jiménez, *et al*., (2006); o como en otros casos, considerar el promedio del primer ciclo estacional como el primer valor inicial de la componente de nivel, valor de cero para la componente de tendencia, y diferencias entre las observaciones y la componente de nivel para determinar la componente estacional; reduciendo de esta forma la cantidad total de datos que pueden aportar información relevante en el modelado.

En este trabajo se utiliza la formulación alternativa en el modelo de pronóstico de Holt-Winters Aditivo trabajado por Bermúdez et. al. (2007) que simplifica la obtención de estimaciones de máximo verosimilitud de los parámetros de suavizado y los valores iniciales, el cálculo de predicciones puntuales e intervalos de predicción fiables. Además, se analiza datos de material particulado (pm2.5) reportados por el Programa Nacional de Vigilancia Sanitaria de Calidad del Aire, Lima–Callao, de la estación DISA, Callao.

Datos de material particulado en el monitoreo de calidad de aire en Lima.

El Material particulado se define como la acumulación de gotitas de un sólido o líquido en la atmósfera ambiental generada a partir de alguna actividad antropogénica o natural (Spiro T., Stigliani W., 1996). En general, el particulado capaz de penetrar las vías respiratorias de los humanos, se divide en 2 rangos de tamaños: (2.5ug/m3 y 10ug/m3). El pm2.5 es responsable de causar los mayores daños a la salud de las personas, siendo hoy objeto de gran atención en los Estados Unidos. El interés en el estudio estadístico del material particulado se centra tanto en las estimaciones puntuales según García (2012); considerando inclusive el horario en que se toman los datos según Ochoa y Jiménez (2011), así como en el análisis de series temporales, con el uso de métodos de suavizado exponencial, según Bedoya y Martínez (2009).

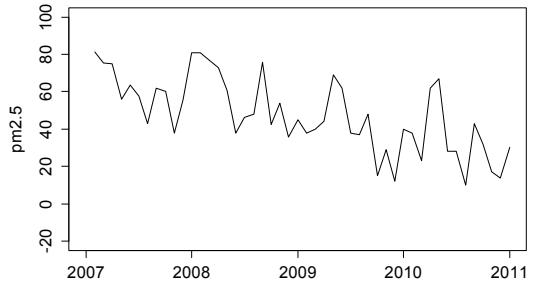

Figura 1. Serie mensual de material particulado (pm2.5), desde Febrero del 2007 hasta Enero del 2011.

Los datos son reportes mensuales, desde febrero del 2007, hasta marzo del 2013, sin embargo, hubo un cese en el monitoreo desde febrero del 2011 hasta marzo del 2012, esto significaba una gran cantidad de datos faltantes, por lo cual se optó solo trabajar con los datos hasta enero del 2011, además de realizar algunas imputaciones en este periodo, siendo 48, la cantidad de datos trabajados. Los datos de marzo y abril del 2008, tienen valores elevados (atípicos), cuya fuente no explica el fenómeno.

Material y métodos

1. Modelo Multivariado de Holt-Winters Aditivo.

El modelo de Holt-Winters Aditivo es usualmente definido a través de las ecuaciones:

$$a_i = \alpha(y_i - c_{i-p}) + (1-\alpha)(a_{i-1} + b_{i-1}) \qquad (ecuación\ de\ nivel)$$
$$b_i = \beta(a_i - a_{i-1}) + (1-\beta)b_{i-1} \qquad (ecuación\ de\ la\ tendencia)$$
$$c_i = \gamma(y_i - a_{i-1} - b_{i-1}) + (1-\gamma)c_{i-p} \qquad (ecuación\ del\ componente\ estacional)$$

donde $Y = (y_i, ..., y_n)$: son los datos observados, p es el tamaño del ciclo estacional y $\theta = (\alpha, \beta, \gamma)'$ es el vector de parámetros de suavización. Además debe considerarse usar un vector de condiciones iniciales $\omega = (a_0, b_0, c_{1-p}, ..., c_0)$. Se asume que $Y_i = a_{i-1} + b_{i-1} + c_{i-p} + \varepsilon_i$, donde el vector de errores $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ sigue una distribución $N_n(0, \sigma^2 I_n)$.

2. El modelo lineal heterocedástico.

Mediante ciertas condiciones, los autores deducen mediante expresiones matriciales el siguiente modelo (Bermúdez et al. 2007):

$$Y = M\psi + L\varepsilon \qquad (1)$$

donde $\psi = (b_0, c_{1-p}, ..., c_0)'$ es el vector de condiciones iniciales, M es una matriz conocida de rango completo $n \times (p+1)$, cuya primera columna está dada por el vector $(0,1,....,n-1)'$ y las siguientes p columnas son construidas por bloques de $p \times p$ matrices identidad, apilados uno sobre otro para cubrir las n filas; definido bajo la restricción de $a_0 + b_0 = 0$, por conveniencia matemática. Por último, la matriz L es una matriz triangular inferior con $l_i = \alpha(1 + (i-1)\beta) + \gamma(i = 1 \bmod p)$ para $i = 2, ..., n$.

3. Estimación de los parámetros.

A partir del modelo mostrado en la ecuación (1), la función de máximo verosimilitud del vector Y es:

$$-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2(\sigma^2)}(Y - M\psi)'(LL')^{-1}(Y - M\psi) \qquad (2)$$

Sea X la matriz $L^{-1}M$, sea $P_x = X(X'X)^{-1}X'$ la matriz de proyección ortogonal en el vector de espacios generados por las columnas de la matriz X, $\tilde{\psi} = (X'X)^{-1}X'L^{-1}Y$ es el estimador cuadrático medio usual de ψ en los modelos heterocedásticos lineales. Finalmente, al reemplazar y descomponer en la expresión (2), la función de máximo verosimilitud es:

$$-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}(\tilde{\psi}-\psi)'X'X(\tilde{\psi}-\psi)-\frac{1}{2\sigma^2}(L^{-1}Y)'(I-P_X)L^{-1}Y \tag{3}$$

de la ecuación (3) el $\hat{\theta}$, estimador de máximo verosimilitud del vector de parámetros de suavizado $\theta=(\alpha,\beta,\gamma)'$ es obtenido, minimizando:

$$\min_{\alpha,\beta,\lambda}\left(L^{-1}Y\right)'\left(I-P_X\right)L^{-1}Y \tag{4}$$

Sea $\hat{L}$ la matriz L calculado en $\hat{\theta}$ y $\hat{X}=\hat{L}^{-1}M$. El estimador de ψ, está dado por $\tilde{\psi}$, es decir:

$$\hat{\psi}=(M'\hat{L}'^{-1}\hat{L}^{-1}M)^{-1}M'\hat{L}'^{-1}\hat{L}^{-1}Y \tag{5}$$

Finalmente el estimador de máximo verosimilitud de σ^2 es:

$$\hat{\sigma}^2=\frac{1}{n}Y'\hat{L}'^{-1}(I-\hat{L}^{-1}M(M'\hat{L}'^{-1}\hat{L}^{-1}M)^{-1}M'\hat{L}'^{-1})\hat{L}^{-1}Y \tag{6}$$

4. Pronóstico.

Ahora, sea Y_1 un vector $n\times1$ de datos observados y Y_2 vector $h\times1$ de datos futuros. Considerando $Y=(Y_1',Y_2')'$ el vector conjunto de $(n+h)\times1$ y suponiendo que aun sigue la distribución presentada en el modelo (1), donde el vector de errores y la matriz M y L son particionados de manera similar al vector Y:

$$Y=\begin{pmatrix}Y_1\\Y_2\end{pmatrix}=\begin{pmatrix}M_1\\M_2\end{pmatrix}\psi+\begin{pmatrix}L_{11}&0\\L_{21}&L_{22}\end{pmatrix}\begin{pmatrix}\varepsilon_1\\\varepsilon_2\end{pmatrix} \tag{7}$$

La distribución condicional de Y_2 dado Y_1 es una Normal Multivariante con media $\mu_{2.1}$ y matriz de varianza $\Sigma_{2.1}$ dado por:

$$\mu_{2.1}=M_2\psi+L_{21}L_1^{-1}(Y_1-M_1\psi)$$
$$\Sigma_{2.1}=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}=\sigma^2L_2L_2'$$

Las predicciones puntuales son dadas por un estimador de la media pronosticada $\mu_{2.1}$. Por tanto la propuesta es usar lo siguiente:

$$\hat{\mu}_{2.1}=M_2\hat{\psi}+\hat{L}_{21}\hat{L}_1^{-1}(Y_1-M_1\hat{\psi}) \tag{8}$$

Si $S=\sigma^{-2}V(Y_2-\hat{\mu}_{2.1})$ y $v\neq0$ es un vector de constantes, entonces:

$$t_v=\sqrt{\frac{n-p-1}{n}}\frac{1}{\hat{\sigma}}\left(v'Sv\right)^{-1/2}v'\left(Y_2-\hat{\mu}_{2.1}\right) \tag{9}$$

Sigue una distribución t-Student con $n-p-1$ grados de libertad, esto permite construir intervalos de predicción exactos para diferentes objetivos, es decir, para un tiempo específico o su acumulativo, dependiendo de la cantidad de constantes (entre unos y ceros) y la posición de ellos en el vector v. Además, siendo usual que el vector θ sea desconocido, el autor propone una aproximación de las estimaciones de los intervalos usando la ecuación (9) con $\hat{\theta}$, aproximando $V\left(Y_2-\mu_{2.1}\right)$ con:

$$\hat{S}=\hat{\sigma}^2(M_2-\hat{L}_{21}\hat{L}_1^{-1}M_1)(M_1'\hat{L}_1'\hat{L}_1'^{-1}M_1)^{-1}(M_2-\hat{L}_{21}\hat{L}_1^{-1}M_1)'+\hat{\sigma}^2\hat{L}_2\hat{L}_2' \tag{10}$$

Resultados

Modelado con la formulación alternativa de Bermúdez *et al.(2007)*

Los resultados presentados fueron calculados mediante el método de Holt-Winters Aditivo con la formulación alterntiva de Bermúdez *et al*. (2007), explicadas en la sección II. Las condiciones iniciales y parámetros de suavizado se obtuvieron minimizando el MSE, en los datos del periodo de ajuste, es decir la maximización de la función de log-verosimilitud (2).

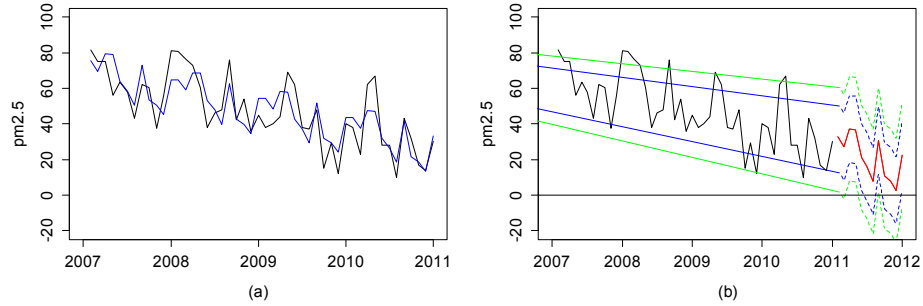

Figura 2. Contraste de datos de pm2.5 observados vs. Ajustados (a). Pronósticos mensual para el periodo Febrero 2011 - Enero 2012, incluido intervalos de confianza al 80% y 95% (b).

En la Figura 2 se muestra los datos ajustados de pm2.5 en el lado (a) (línea azul) en contraste con los datos observados (línea negra), que presentan cierta estacionalidad y tendencia negativa. Los pronósticos de los siguientes 12 meses (línea roja) son mostrados en el lado (b), junto a las estimaciones de los intervalos de predicción al 80% (Hansen, 2012) mostradas en línea segmentada azul y al 95% en línea segmentada verde. Los intervalos de predicción indican posibles valores negativos en las predicciones, por tanto las observaciones se transformaron a logaritmos para restringir valores por debajo de cero.

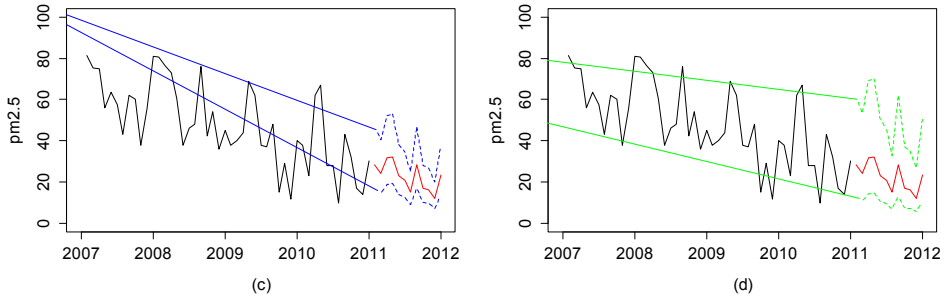

Figura 3. Pronóstico con transformación"log", para mejorar el análisis. Intervalos de predicción al 80% en línea segmentada azul (c). Intervalos de predicción al 95% en líneasegmentada verde (d).

Las condiciones iniciales son mostradas en la Tabla 1: nivel, tendencia y parámetros de suavizado; y en la Tabla 2:condiciones iniciales del componente estacional. Según estas estimaciones, los parámetros de suavizado deben ser ceros, y el modelo correspondiente es lineal con un valor de nivel inicial de 4.307, con tendencia negativa de -0.023 y una componente estacional aditiva(p = 12).

Tabla 1. Componentes de nivel, tendencia y parámetros de suavización iniciales estimados.

Nivel	Tendencia	Alfa	Beta	Gamma
4.307	-0.023	0.000	0.000	0.000

Tabla 2. Componentes de estacionalidad iniciales estimados.

Componentes	Enero	Febrero	Marzo	Abril	Mayo	Junio
Valores	0.138	0.014	0.296	0.331	0.023	-0.052
Componentes	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
Valores	-0.357	0.301	-0.191	-0.220	-0.475	0.194

Tabla 3. Pronóstico e intervalos de predicción al 80% y 95% de pm2.5 en Lima - Callao, para 12 meses.

		Al 80%		Al 95%	
Año-mes	Pronóstico	Límite inferior	Límite superior	Límite inferior	Límite superior
2012-Feb	28.247	17.064	46.760	12.905	61.828
2012-Mar	24.398	14.738	40.388	11.147	53.402
2012-Abr	31.633	19.109	52.364	14.452	69.238
2012-May	32.025	19.346	53.013	14.631	70.097
2012-Jun	23.013	13.902	38.096	10.514	50.372
2012-Jul	20.878	12.612	34.561	9.539	45.698
2012-Ago	15.041	9.086	24.899	6.872	32.922
2012-Sep	28.406	17.160	47.023	12.978	62.176
2012-Oct	16.987	10.262	28.120	7.761	37.181
2012-Nov	16.138	9.749	26.714	7.373	35.322
2012-Dic	12.223	7.384	20.233	5.584	26.754
2013-Ene	23.318	14.086	38.600	10.653	51.038

Resultados con métodos automáticos de Hyndman*et al*. (2008)

En esta sección se hizo uso de los métodos automáticos creados por Hyndman *et. al* (2008), en la librería "`forecast`" en R. Se calcula el modelo y pronóstico tanto para los modelos de Holt-Winters, como para los modelos ARIMA, y se comparan los parámetros inciales junto con las medidas de bondad de ajuste. El modelo automático de Holt-Winters estima un componente de nivel es multiplicativo, no tiene tendencia y la componente estacional es aditiva; además, se presentan los intervalos de predicción al 80% de confianza (Figura 4e.), y 90% de confianza (Figura 4f.). El modelo automático de ARIMA, determina un modelo autorregresivo de orden uno, con primeras diferencias y componente estacional. Sus intervalos de predicción al 80% y 95% de confianza se muestran en la Figura 5(g) y 5(h).

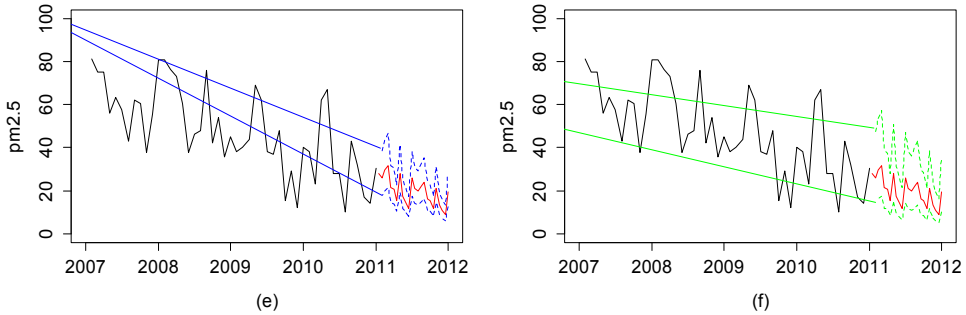

Figura 4. Modelo Holt-Winters, Hyndman*et al*.(2008).Pronóstico de 12 meses e intervalos de predicción al 80% (e) y 95% (f).

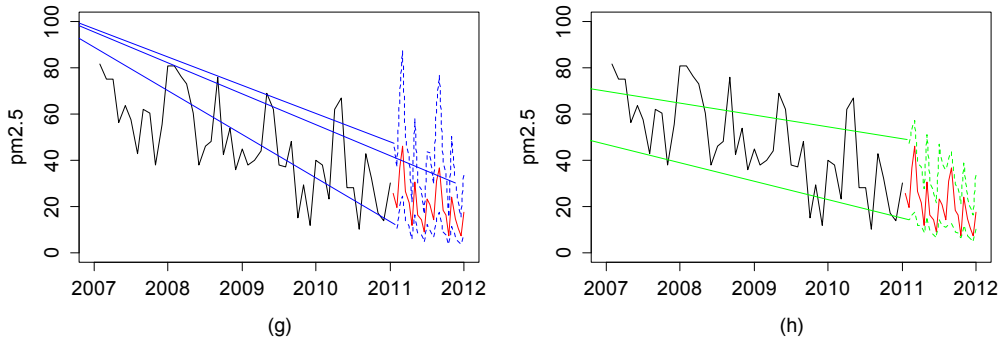

Figura 5. Modelo ARIMA (1,1,0)(12), Hyndman*et al*. (2008). Pronóstico de 12 meses e intervalos de predicción al 80%(g) y 95%(h).

Dado que los pronósticos puntuales y los intervalos podrían resultar valores negativos, fue conveniente trabajar con los logaritmos de los datos, luego estos resultados retornados a su escala original para su interpretación adecuada. Los pronósticos se realizan para 12 meses de manera puntual y con intervalos de predicción, que se muestran en la Tabla 3. El modelo de Holt-Winters por Bermúdez *et al*.(2007), muestra valores de bondad de ajuste muy similares a los métodos automáticos Holt-Winters y ARIMA. El modelo de Holt-Winters automático, considera parámetros de suavizado en estacionalidad (gama = 0.0806) y muestran mayor error en términos generales de bondad de ajuste respecto al modelo de Bermúdez *et al.*(2007), que considera a los tres parámetros como valores cero (Tabla 4).

Tabla 4. Comparativa de condiciones iniciales y parámetros de suavizado.

Modelos	Alfa	Beta	Gama	a_0	b_0
Holt-Winters (Bermudez *et al*.)	0.0000	0.0000	0.0000	4.307	-0.023
Holt-Winters (Hyndman*et al.*)	0.0003	0.0001	0.0806	4.3202	-0.023

Tabla 5. Comparativa de medidas de bondad de ajuste.

Medidas de ajuste.	H-W.Bermúdez et al.	H-W.Automático	ARIMA(1, 1, 0)(12)
RMSE	12.626	12.973	12.802
MAD	9.647	9.869	8.988
MAPE	23.056	24.307	25.747
SMAPE	21.737	22.745	22.934

Conclusiones

El modelo de Holt-Winters, deducido por Bermúdez *et al* (2007), genera estadísticos muy similares respecto los métodos automáticos de Hyndman *et al* (2008), tanto en modelos Holt-Winters y ARIMA. La bondad de ajuste de los tres modelos son similares, por tanto, la decisión se basa en la viabilidad de los resultados. Los pronósticos deben tener valores no muy cercanos a cero, al igual que sus intervalos de predicción.

Agradecimiento

Este trabajo fue realizado gracias al financiamiento de la Oficina de Relaciones Internacionales y Cooperación de la Universitat de Valencia mediante la "Beca de jóvenes investigadores de países en vías de desarrollo 2013"; bajo la supervisión del Dr. José D. Bermúdez Edo, profesor titular del Departamento de Estadística e Investigación Operativa.

Referencias bibliográficas

Bedoya, J., & Martínez, E. (2009). Calidad del aire en el Valle de Aburrá Antioquia-Colombia. *Dyna* , 7-15.

Bermúdez, J., Vercher, E., & Segura, J. (2007). Holt-Winters forecasting: an alternative formulation applied to UK air passenger data. *Journal Applied Statistics*, 34 (9), 1075-1090.

Chatfield, C., &Yar, M. (1991). Prediction intervals for multiplicative Holt–Winters. *Int J Forecast* , 7, 31-37.

García, E. J. (2012). Estadística bayesiana y su naturaleza continúa en la estimación de niveles contaminantes por material particulado pm10 en control y monitoreo de calidad del aire en Trujillo (La Libertad). *Conocimiento para el desarrollo*, 3 (1), 93-100.

Gardner, J. (2006). Exponential smoothing: The state of the art-part II. *Int J Forecast* (22), 637-666.

Hyndman, R., Koehler, A., Ord, K., & Snyder, R. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin: Springer.

Jiménez, J. F., Gázquez, J. C., & Sánchez, R. (2006). La capacidad predictiva en los métodos Box-Jenkins y Holt-Winters: una aplicación al sector público. *Revista Europea de Dirección y Economía de la Empresa*, 15 (3), 185-192.

Ochoa, A., & Jiménez, J. (2011). *Ciclo diurno de PM10 en el Valle de Aburra*. IX Congreso Colombiano de Meteorología. Bogotá D.C.

Ord, J. K., Koehler, A. B., & Snyder, R. D. (1997).Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models. *Journal of the American Statistical Association* (92), 1621-1629.

R Core Team (2013). R: *A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/.

Spiro, T. G., & Stigliani, W. M. (1996). *Química Medioambiental*. Madrid: Pearson Prentice Hall.