MODELACIÓN DE TEMPERATURAS DEL AIRE A 850 MB EN EL CASO DE LA ONDA CÁLIDA DE JULIO DEL 2006 EN EL NOROESTE DE MÉXICO

Elvia CONTRERAS-NAVARRO[1], O. Rafael GARCIA CUETO[2], Félix F. GONZALEZ-NAVARRO[3], Juan Ramón CASTRO-RODRÍGUEZ[4]
[1,2,3] *Instituto de Ingeniería, Universidad Autónoma de Baja California, México*
[4] *Facultad de Ingeniería, Universidad Autónoma de Baja California, México*
elvia.contreras@uabc.edu.mx, rafaelcueto@uabc.edu.mx, fernando.gonzalez@uabc.edu.mx, jrcastror@uabc.edu.mx.

Resumen
El estudio de las ondas de calor se enfrenta al reto de conocer los elementos clave que intervienen en su formación y desarrollo y la manera en que afectan a la salud de la población, lo que en un futuro permitiría su modelación predictiva. Se analiza el caso de una onda cálida que afectó a la ciudad de Mexicali, Baja California, México, en los días del 16 al 26 de julio de 2006 que causó graves estragos en la población. Se identifican las principales variables que causan el desarrollo de una onda cálida, y se propone la modelación de la temperatura del aire a 850 mb ($Taire_{850}$) mediante un modelo de regresión múltiple (MRM) y regresión con redes neuronales (RRN). Se encuentra que el MRM explica a la $Taire_{850}$ un 67%, siendo la altura geopotencial a 500mb y la presión atmosférica al nivel del mar las variables que más contribuyen a esta explicación. En cuanto a la RRN muestra un desempeño más aceptable i.e. coeficientes de regresión de 0.89; el p-value obtenido fue de 0.0019, por lo que se puede afirmar que un modelo no lineal como la Red Neuronal tiene un mejor desempeño en la predicción de valores independientes que un modelo lineal.

Palabras clave: Onda cálida, temperatura del aire, regresión múltiple, redes neuronales.
.

Abstract
The study of heat waves is challenged to know the key elements involved in their formation and development, and how they affect the health of the population, which in the future would allow its predictive modeling. The case of a heat wave that affected the city of Mexicali, Baja California, Mexico, on days 16 to July 26, 2006 causing serious damage on the population is analyzed. Major variables that cause the development of a heat wave are identified, and the modeling of air temperature at 850 mb ($Taire_{850}$) is proposed by a multiple regression model (MRM) and regression neural network (RNN). It was found that the MRM explained to $Taire_{850}$ a 67%, with a 500mb geopotential height and atmospheric pressure at sea level the atmospheric variables that more contribute to this explanation. Insomuch as RNN show a performance more, i.e. regression coefficients of 0.89; the p-value obtained was 0.0019, so we can say that a nonlinear model such as neural network has a better performance in predicting independent values that a linear model.

Key words: Heat wave, air temperature, multiple regression, neural network

1. INTRODUCCION

Las altas temperaturas del aire pueden afectar la salud humana y exacerbar condiciones de morbilidad en las poblaciones afectadas. Los grupos de población que tienen un elevado riesgo ante eventos cálidos extremos son los más ancianos, los niños, o quienes tienen problemas físicos o mentales. Los cambios de magnitud y frecuencia de las ondas de calor han ocasionado grandes impactos sobre la salud de las poblaciones. Estudios recientes (Tobías et al, 2009; Haines et al ,2005; Kysel, 2004) han analizado los efectos en la salud derivados de la exposición durante largo tiempo a altas temperaturas. Tal es el caso de los denominados golpes de calor, que se trata de la alteración más grave de la regulación térmica corporal (ocurre cuando la temperatura corporal rebasa los 40°C) y cuya relación son los incrementos de morbilidad y mortalidad.
En particular el noroeste de México es especialmente vulnerable a los inminentes cambios del clima mundial y regional. De acuerdo a los escenarios que presenta el IPCC, la región noroeste de México tendrá una disminución del 10 al 20% en su precipitación total anual, mientras que la temperatura media anual aumentará entre 1.5 y 2.5°C en los próximos 50 años. En el estado de Baja California ya hay estudios que evidencian alteraciones a eventos extremos; se ha observado un aumento en las ondas cálidas (García *et al*., 2010) con un impacto fuerte en el sector salud; de hecho en los últimos cuatro años (2004-2007) han ocurrido 43 defunciones por el llamado golpe de calor en el municipio de Mexicali. Por lo tanto, es de suma importancia analizar con detalle los factores que producen las ondas de calor para modelar su comportamiento y en un futuro proponer un sistema de alerta temprana para este evento atmosférico extremo. En función de lo anterior es que en este artículo se propone avanzar en ese conocimiento con la propuesta de la modelación de la temperatura del aire a 850 milibares en la onda cálida ocurrida en el mes de julio del año 2006. Nuestra hipótesis es que si logramos encontrar una buena explicación del comportamiento de la temperatura a ese nivel, también lo tendremos con las temperaturas que se presentan a nivel de superficie, y con ello la selección de variables adecuadas para explicar el comportamiento de las ondas cálidas.
En este sentido, son dos los enfoques presentados en este trabajo, por un lado se utiliza un modelo de regresión múltiple desde un punto de vista de la estadística clásica; y por el otro, se utiliza el enfoque de aprendizaje de máquina en el que dos algoritmos de aprendizaje, como son las redes neuronales y mínimos cuadrados parciales, son entrenados a fin de generar modelos de predicción.

2. METODOLOGÍA

2.1 PREDICCION CON REDES NEURONALES

2.1.1 REGRESION CON REDES NEURONALES

Las redes neuronales pueden ser entendidas como modelos de regresión. Por lo tanto se emplean en muchas ocasiones como herramienta para predecir valores futuros de una o múltiples variables objetivo, que en estadística son variables de respuesta. Muchos métodos estadísticos clásicos y otros más recientes han sido reescritos, no siempre de forma consciente, como redes neuronales. Esto nos da idea de lo generales que pueden llegar a ser las estructuras representadas a través de un esquema de redes neuronales, y de su clara relación con la estadística (Castellano, 2009).

2.1.2 ALGORITMO DE MÍNIMOS CUADRADOS PARCIALES

El método de mínimos cuadrados parciales (PLS) es una extensión del modelo de regresión lineal múltiple y se encuentra relacionado con el análisis de componentes principales (PCA). En su forma más simple un modelo lineal especifica la relación (lineal) entre la variable dependiente Y, y un grupo de variables predictivas X (Gonzalez y Alciaturi, 2012).

En este paso se utilizó el programa del algoritmo de mínimos cuadrados parciales Este programa permite obtener el vector de regresión óptimo con los datos de referencia para luego validarlo con los datos de validación. Este programa permite encontrar el número de variables latentes que están más correlacionadas con la variable respuesta. El número óptimo de variables latentes corresponde al que arroje el menor valor en la suma de los cuadrados del error general.

2.2 VALIDACION CRUZADA.

Validación cruzada es un método de evaluación del modelo que sea mejor que los residuales. El problema con las evaluaciones residuales es que no dan una indicación de lo bien que va a hacer nuevas predicciones para los datos que ya no se ha visto. Una forma de superar este problema es no utilizar todo el conjunto de datos al entrenar a un aprendiz. Algunos de los datos son eliminados antes de que comience el entrenamiento. Luego, cuando se realiza la formación, los datos que se elimina se puede utilizar para probar el rendimiento del modelo aprendido en `` nuevos" datos. Esta es la idea básica para toda una clase de métodos de evaluación modelo llamado validación cruzada.
El método de retención es el tipo más simple de validación cruzada. El conjunto de datos se divide en dos grupos, llamado el conjunto de entrenamiento y el conjunto de pruebas. El aproximador funcional se ajusta a una función utilizando el conjunto de entrenamiento solamente. Entonces el aproximador funcional se preguntó para predecir los valores de salida de los datos en el conjunto de pruebas (que nunca ha visto a estos valores de salida antes). Los errores que hace se acumulan como antes para dar el error medio de prueba absoluta, que se utiliza para evaluar el modelo. La ventaja de este método es que por lo general es preferible el método residual y ya no se necesita para calcular. Sin embargo, su evaluación puede tener una alta varianza. La evaluación puede depender en gran medida de que los puntos de datos terminan en el conjunto de entrenamiento y que terminan en la prueba, y por lo tanto la evaluación pueden ser muy diferentes en función de cómo se hace la división.

K veces la validación cruzada es una forma de mejorar el método de exclusión. El conjunto de datos se divide en k subconjuntos, y el método de retención se repite k veces. Cada vez, uno de los subconjuntos de k se utiliza como equipo de prueba y los otros k-1 subconjuntos se juntan para formar un conjunto de entrenamiento. Luego se calcula el error promedio en todos los ensayos k. La ventaja de este método es que importa menos cómo los datos se dividen. Cada punto de datos llega a ser en un montaje de prueba exactamente una vez, y llega a ser en un conjunto de entrenamiento k-1 veces. La varianza de la estimación resultante se reduce a medida que aumenta k. La desventaja de este método es que el algoritmo de entrenamiento tiene que volver a ejecutar desde tiempos scratch k, lo que significa que se necesita k veces tanto la computación para hacer una evaluación. Una variante de este método consiste en dividir aleatoriamente los datos en una prueba y la formación conjunto K diferentes momentos.

La ventaja de hacer esto es que usted puede elegir de forma independiente la extensión de cada conjunto de pruebas y cuántas pruebas que lo habitual sobre.

2.3 DESCRIPCION DE DATOS

Las variables meteorológicas más importantes que parecen intervenir en el desarrollo de una onda cálida en el norte de Baja California, y que serán incorporadas en este análisis exploratorio son: 1) las temperaturas del aire a 850 milibares, 2) Altura geopotencial (m) a 500 milibares (*Alt Geop*$_{500}$), 3) Componente u del viento a 500 milibares (*Vto U*$_{500}$), 4) Componente v del viento a 500 milibares (*Vto V*$_{500}$), y 5) Presión atmosférica en milibares al nivel del mar (*Patm*). En este caso la modelación de la temperatura del aire a 850 milibares se puso en función de las otras cuatro variables restantes. El caso de estudio es la onda cálida de julio de 2006, y lo primero que se hizo fue la recolección de datos climatológicos, lo cual se realizó de la siguiente manera:

- Se identificaron las variables que intervienen en el fenómeno de la onda de calor, las cuales ya fueron enunciadas en el párrafo anterior.
- Los datos diarios para cada variable fueron descargados directamente en la página de la NOAA (http://www.esrl.noaa.gov/) del día 15 al 31 de Julio del 2006.
- Se utilizo una malla de 30° a 45° latitud norte y longitud 125° a 100° oeste. Las herramientas utilizadas para procesar los datos fueron Python con las librerías: Numpy y NetCDF.

De esta manera, se dispone de un conjunto de datos de 1309 observaciones con cuatro predictores y una variable de respuesta. Posteriormente se extrajo de forma aleatoria el 30% de las observaciones, a fin de generar un conjunto de datos de prueba independiente y poder verificar el rendimiento de los modelos de predicción.

2.4 MODELOS DE PREDICCIÓN Y VALIDACIÓN

Con el 70% de los datos se indujeron dos modelos de predicción ampliamente usados, Mínimos Cuadrados Parciales y una Red Neuronal para regresión. Ambos modelos fueron entrenados mediante validación cruzada en su modalidad de 10 repeticiones por 10 divisiones. La medida de evaluación de la validación cruzada fue el NRMSE *Normalized Root Mean Square Error*, la cual de define de la siguiente manera:

$$NRMSE = \sqrt{\frac{1/n \sum_{i=1}^{n} (observed - predicted)^2}{1/n \sum_{i=1}^{n} (observed - \overline{observed})^2}}$$

El modelo de Red Neuronal es del tipo *Feedforward-Backpropagation*, teniendo una topología de dos capas internas con cinco neuronas cada una de ellas. La estrategia de aprendizaje de la red es el algoritmo *de Levenberg–Marquardt* el cual emplea una aproximación a la matriz Hessiana mediante el cuadrado de la matriz Jacobiana. Este modelo tiene la ventaja de no tener parámetros de ajuste, ya que el gradiente de actualización de pesos es derivado del Jacobiano[1] de la función de error y el vector de

[1] Sea F(w) la función de error con respecto a los pesos. El Jacobiano se define como

$$\nabla F(w) = \left[\frac{\partial F(w)}{\partial w_1} \cdots \frac{\partial F(w)}{\partial w_n} \right]$$

errores de predicción. Una vez entrenados los modelos, se utilizaron para hacer pruebas de predicción en el conjunto independiente.

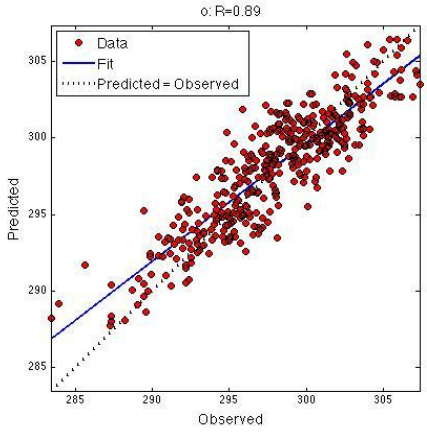

Figura 1: ANÁLISIS DE REGRESIÓN SIMPLE DE LA PREDICCIÓN GENERADA POR LA RED NEURONAL VS. LOS DATOS OBSERVADOS EN LA MUESTRA INDEPENDIENTE DE PRUEBA.

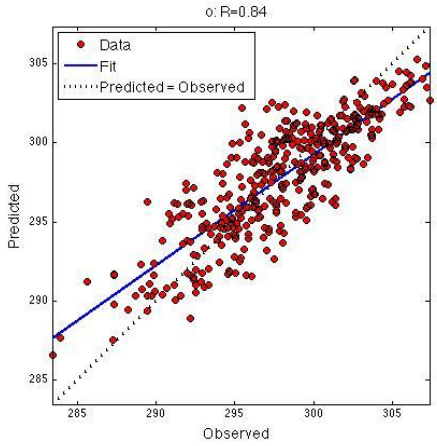

Figura 2: ANÁLISIS DE REGRESIÓN SIMPLE DE LA PREDICCIÓN GENERADA EL MODELO DE MÍNIMOS CUADRADOS PARCIALES VS. LOS DATOS OBSERVADOS EN LA MUESTRA INDEPENDIENTE DE PRUEBA.

Medidas estadística	Mínimos Cuadrados Parciales	Red Neuronal
NRMSE	0.58	0.54
Error estándar	0.01	0.01
Límite inferior	0.57	0.53
Límite superior	0.59	0.56

Tabla 1: RESULTADOS DEL PROCESO DE ENTRENAMIENTO DE MODELOS POR VALIDACIÓN CRUZADA. NRMSE=NORMALIZED ROOT MEAN SQUARE ERROR.

2.5 MODELO ESTADISTICO

En este análisis exploratorio se propone un modelo de regresión múltiple para ver la relación que guarda la primera de esas variables, temperatura del aire a 850 milibares, que de aquí en adelante se abreviará como $Taire_{850}$, con las cuatro variables restantes.
En la propuesta del modelo de regresión múltiple se utilizó el software Statistica, y se realizó una regresión por pasos (stepwise) para ver la evolución del coeficiente de determinación y su nivel de significancia asociado. Se realizaron las correlaciones teniendo como variable dependiente a $Taire_{850}$, y como variables independientes a $Alt\ Geop_{500}$, $Patm$, $Vto\ U_{500}$ y $Vto\ V_{500}$.

3. RESULTADOS

3.1 MODELO DE REGRESION MULTIPLE

La matriz de correlaciones de las cinco variables mencionadas en la metodología se visualiza en la fig. X; se puede ver que las variables $Alt\ Geop_{500}$ y $Patm$ parecen estar mejor correlacionadas con la $Taire_{850}$.

Variables independientes	Paso +dentro/-fuera	R Múltiple	R^2 Multiple	Cambio en R^2	F	Nivel p	Variables incluidas
$Alt\ Geop_{500}$	1	0.680364	0.462895	0.462895	1126.417	0.000000	1
$Patm$	2	0.807829	0.652588	0.189693	713.097	0.000000	2
Vto	3	0.819247	0.671165	0.018577	73.725	0.000000	3
Vto	4	0.822284	0.676151	0.004986	20.077	0.000008	4

Tabla 2 RESUMEN DE REGRESIÓN POR PASOS (STEPWISE). LA VARIABLE DEPENDIENTE ES $Taire_{850}$.

En la tabla 2 se muestra el resumen de la regresión múltiple para la variable dependiente $Taire_{850}$. Se presentan los coeficientes de regresión estandarizados (Beta) y sus errores estándar, los estimadores crudos de regresión (B) y sus errores estándar asociados, así como la prueba t, con 1304 grados de libertad, y el nivel estadístico de significancia p. En particular, los errores estándar de los coeficientes estandarizados (Beta) y crudos (B) de las variables independientes son pequeños, comparados con los valores de los mismos coeficientes, lo que nos permite asegurar que son casi ciertamente diferentes de cero. La correlación (R) tiene un valor de 0.822, y el coeficiente de determinación de 0.676. El valor del estadístico F es muy alto (680.64), lo que fortalece la relación lineal propuesta, aunado al nivel de significancia estadística asociado ($p < 0.0000$). El error estándar del estimador ($s_e = 2.52$) nos permite hablar de la precisión de estimar las $Taire_{850}$, en base a las cuatro variables predictoras seleccionadas, ya que uno espera que alrededor del 95% de los valores de la $Taire_{850}$ estén dentro de $\pm 2s_e = 5.0°C$ de las temperaturas estimadas por el modelo de regresión.

El modelo de regresión múltiple, elegido en el paso 4 del método por pasos (stepwise) de acuerdo a los parámetros de regresión (B) estimados con el conjunto de datos utilizado es:

$$Taire_{850} = 320.75 + 0.0601 * Alt\ Geop_{500} - 0.372 * Patm - 0.100 * Vto + 0.0701 * Vto$$

Regresión Múltiple para la Variable Dependiente $Taire_{850}$
R= .822; R^2= .676; R^2 ajustado= .675;F(4,1304)=680.64, p<0.0000,
Error estándar del estimador: 2.52

	Beta	Error Estándar de Beta	B	Error Estándar de B	t (1304)	Nivel de significancia p
Intercepto			320.75	17.13	18.71	0.000000
$Alt\ Geop_{500}$	0.614	0.0181	0.0601	0.0017	33.77	0.000000
Patm	-0.407	0.0161	-0.372	0.0148	-25.14	0.000000
$Vto\ U_{500}$	-0.166	0.0185	-0.100	0.0111	-8.98	0.000000
$Vto\ V_{500}$	0.0718	0.0160	0.0701	0.0156	4.48	0.000008

Tabla 3 RESUMEN DE REGRESIÓN MÚLTIPLE PARA LA VARIABLE DEPENDIENTE $Taire_{850}$

En base a lo analizado se resume que las variables independientes más importantes para explicar a la $Taire_{850}$ son la $Alt\ Geop_{500}$ y la *Patm*; las medidas de bondad de ajuste, como el R^2, la prueba F y el error estándar de los estimadores, confirman lo anterior. Las componentes del viento a 500 milibares no parecen ser variables que deban ser introducidas en el modelo final de esta investigación preliminar. Se debe comentar que la propuesta de este modelo no es más que un paso inicial que permitirá avanzar en la elección de las variables más importantes que tienen que ver con la génesis y desarrollo de las ondas cálidas en el noroeste de México.

3.2 MODELO DE REDES NEURONALES

Los errores en la validación arrojan valores bajos y una variabilidad en las lecturas consistentemente bajos i.e. un error estándar de 0.01 para los dos modelos. Por otra parte el análisis de regresión simple entre las observaciones y las predicciones de ambos modelos, muestra un desempeño aceptable i.e. coeficientes de regresión 0.89 en la Red Neuronal. El modelo lineal generado por el método de Mínimos Cuadrados Parciales, tiene un desempeño un tanto similar que la Red pero un poco menor. Ante esto, el test de suma de rankeo de Wilcoxon de dos lados fue aplicado para fijar esta diferencia de desempeño. El p-value obtenido fue de 0.0019, por lo que se puede afirmar que para el problema en particular, un modelo no lineal como la Red Neuronal tiene un mejor desempeño en la predicción de valores independientes que un modelo lineal, como el Mínimos Cuadrados Parciales.

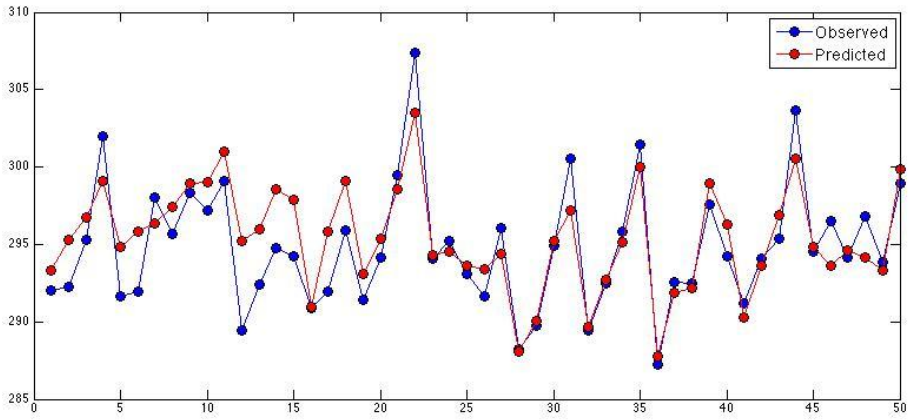

Figura 3: VENTANA DE 50 PREDICCIONES DE LA RED NEURONAL VS DATOS OBSERVADOS EN LA MUESTRA INDEPENDIENTE DE PRUEBA.

En la Figura 3, se extrajo una ventana de 50 puntos de temperatura para observar gráficamente el desempeño del mejor modelo. En ésta se puede observar que la Red Neuronal *sigue* la distribución de temperaturas en ese rango de puntos en particular, de una manera aceptable.

4. CONCLUSIONES

En este estudio se identificaron las principales variables que causan el desarrollo de una onda cálida, y se propone la modelación de la temperatura del aire a 850 mb ($Taire_{850}$) mediante un modelo de regresión múltiple (MRM) y regresión con redes neuronales (RRN). Se encontró que el MRM explica a la $Taire_{850}$ un 67%, siendo la altura geopotencial a 500mb y la presión atmosférica al nivel del mar las variables que más contribuyeron a esta explicación. Las Redes Neuronales mostraron un desempeño más acceptable, ya que se obtuvieron coeficientes de regresión de 0.89; la significancia estadística obtenida fue de 0.0019, por lo que se puede afirmar que un modelo no lineal como la Red Neuronal tiene un mejor desempeño en la predicción de valores independientes que un modelo lineal. El siguiente paso en este esquema de modelación será estudiar más casos de ondas cálidas y el de analizar las temperaturas de superficie locales asociadas con estos eventos extremos.

5. BIBLIOGRAFIA

Castellano, M. M. (2009). Modelización estadística con redes neuronales. Aplicaciones a la hidrología, aerobiologia y modelización de procesos. Tesis doctoral, Universidad de Da Coruña.

Consultada (2014). School of computer science http://www.cs.cmu.edu/~schneide/tut5/node42.html.

García, C.O.R., Tejeda, M.A., Jáuregui, E.,(2010). Heatwaves and heatdays in an arid city in the northwest of Mexico: current trends and in climate change scenarios. International J. Biometeorol., 36-46 pp.

González, P.,_Alciaturi, C., (2012). Desarrollo De Un Programa Para Estudiar El Comportamiento De Una Columna De Fraccionamiento Etano/Etileno De Una Planta De Olefinas. (Development Of A Program For Studying The Behavior Of An Ethane - Ethylene Splitter Of An Olefins Plant)

Haines A., Kovats R.S., Campbell Lendrumb D, Corvalan C., (2006). Climate change and human health: Impacts, vulnerability and public health. Public Health vol 120, 585–596pp.

Kysel J.,, (2004). Mortality And Displaced Mortality During Heat Waves In The Czech Republic. Int J Biometeorol 49:91–97 pp.

Tobías A., García de Olalla P , Linares C., Bleda M.. Caylà J,, Díaz J. (2009). Short-term effects of extreme hot summer temperatures on total daily mortality in Barcelona, Spain. Int J Biometeorol