Posgrado de Ingeniería de Sistemas

Convolución y Chi cuadrada

Alberto Benavides
17 de noviembre de 2020

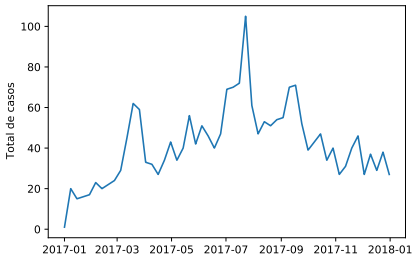
1. Sobre los datos

Los datos para esta tarea provienen del tema de tesis que actualmente trabajo: relación entre contaminantes y enfermedades en el *área metropolitana de Monterrey* (AMM). Los datos de los contaminantes fueron obtenidos del SIMA [3], mientras que los datos de las enfermedades provienen de la página de la Secretaría de Salud de México [2]. Los datos de contaminantes fueron tomados cada segundo por las distintas trece estaciones de monitoreo ambiental ubicadas en el AAM e incluyen concentraciones de algunos contaminantes, de los que se destacan los que tienen tamaños de partículas de 10 PM medidas en μg / m^3, en tanto que los datos de las enfermedades contienen información sobre la edad, género, talla, enfermedad y, dato de principal interés para esta tarea, la fecha en la que se dio la consulta de cada caso.

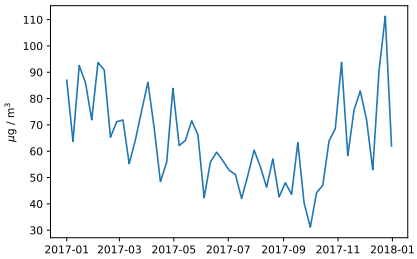
2. Preprocesamiento de los datos

A ambos conjuntos de datos se les eliminaron las inconsistencias como valores fuera de rangos permitidos o la ausencia de valores, posteriormente se obtuvieron sólo concentraciones de contaminantes y consultas del año 2017 por ser el año en que se hallan datos más completos. Después, se agruparon estos datos semanalmente, obteniendo el promedio semanal de las concentraciones de partículas de 10 PM y la suma de todos los casos de consultas registradas, ambos datos mostrados en la figura 1 (p. 2).

Una práctica común para comparar series de tiempo es normalizarlas y diferenciarlas mediante la resta de cada valor, menos su valor en una unidad de tiempo anterior, lo que hace que la serie de tiempo resultante tenga valores en torno a cero entre $[-1, 1]$, como puede verse en la figura 2 (p. 2).



(a) Casos.



(b) Partículas de 10 PM.

Figura 1: Series de tiempo de casos registrados de enfermedades y concentraciones de partículas de 10 PM en el AMM durante 2017.

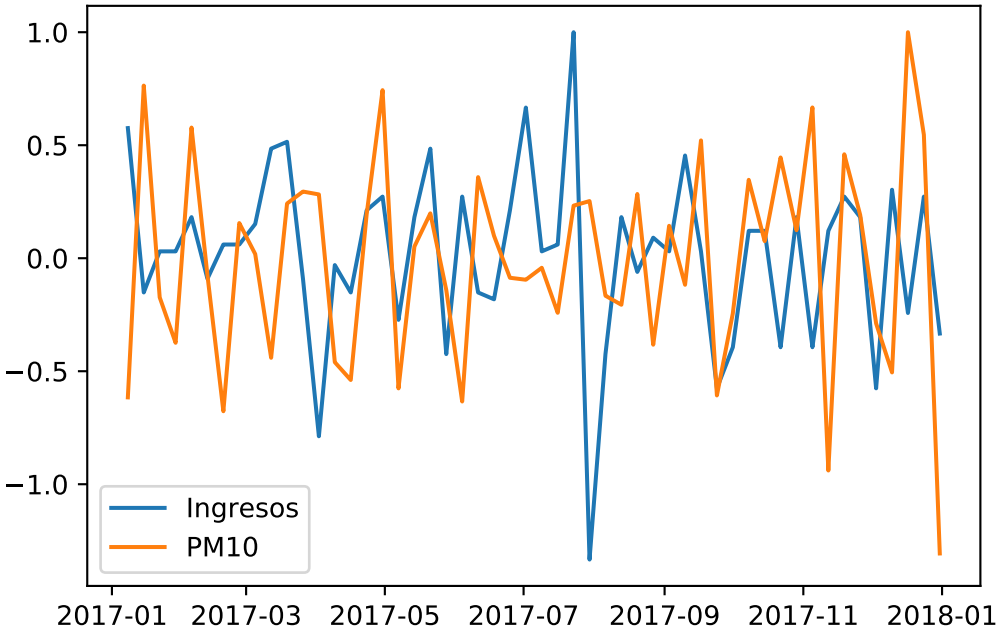


Figura 2: Series de tiempo diferenciadas de casos registrados de enfermedades y concentraciones de partículas de 10 PM en el AMM durante 2017.

3. Chi cuadrada

La prueba de Chi cuadrada tiene dos usos principales: Determinar la independencia entre conjuntos de datos y conocer si valores observados son similares a valores esperados (a esto último se le llama prueba de ajuste). En este caso, se utiliza la prueba de Chi cuadrada para determinar si los valores diferenciados y normalizados (como lo sugieren Hyndman y Athanasopoulos [1]) de los contaminantes son independientes de los de las consultas de las enfermedades registradas y, también, si los casos registrados observados son similares a los esperados, tomados estos últimos como las concentraciones en promedio de las partículas de 10 PM.

Para el caso de la independencia, la hipótesis nula es que ambas variables son independientes. El p-valor obtenido es 1 por lo que se acepta la hipótesis nula. Mientras que para la prueba de ajuste, la hipótesis nula es que los valores observados y esperados no tienen diferencia, pero ésta hipótesis se rechaza dado que el p-valor $= 6.07 \times 10^{-175}$.

4. Convolución

Una convolución muestra la distribución de probabilidad Z de la suma de dos variables aleatorias X y Y tal que $P(Z = j) = \sum_{i=-\infty}^{\infty} P(X = i) \times P(Y = j - i)$. Para el caso discreto, esto se puede calcular como $f_c(i) = \sum_j f_1(j) \times f_2(i-j)$, en tanto para el continuo se tiene $(f * g)(z) = \int_{-\infty}^{\infty} f(z - x) \times g(x)\mathrm{d}x$. Este concepto de convolución también se puede utilizar para conocer las interacciones entre señales o, en este caso, series de tiempo de modo que se obtiene el área común a ambas series de tiempo, es decir, el grado de relación de una (fija) respecto a la otra (que se desplaza en el tiempo). La convolución de las series de tiempo diferenciadas puede verse en la figura 3 (p. 4).

5. Propiedades de la varianza

Existen dos propiedades de la varianza, a saber $\mathrm{Cov}[aX + b, cY + d] = ac\mathrm{Cov}[X, Y]$ y $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X, Y]$ que pueden demostrarse numéricamente al generar dos conjuntos de variables aleatorias $X \sim \mathcal{U}(0, 1)$ y $Y \sim \mathcal{B}(100, 0.05)$ de mil valores cada uno. Los histogramas de estos conjuntos aparecen en la figura 4 (p. 5). Para comprobar esto, se toman enteros aleatorios entre $[-10, 10]$ para a, b, c, d y mediante las funciones `var` y `cov` de R, que se utilizan para calcular la varianza y covarianza en ese orden, se comprueba que ambas propiedades se cumplen, como se muestra en el código 1

Código 1: Demostración numérica de propiedades de varianza y covarianza

```
1  X = runif(1000)
2  Y = rbinom(1000, 100, 0.05)
3
4  a = sample(-10:10, 1)
5  b = sample(-10:10, 1)
6  c = sample(-10:10, 1)
7  d = sample(-10:10, 1)
8
```

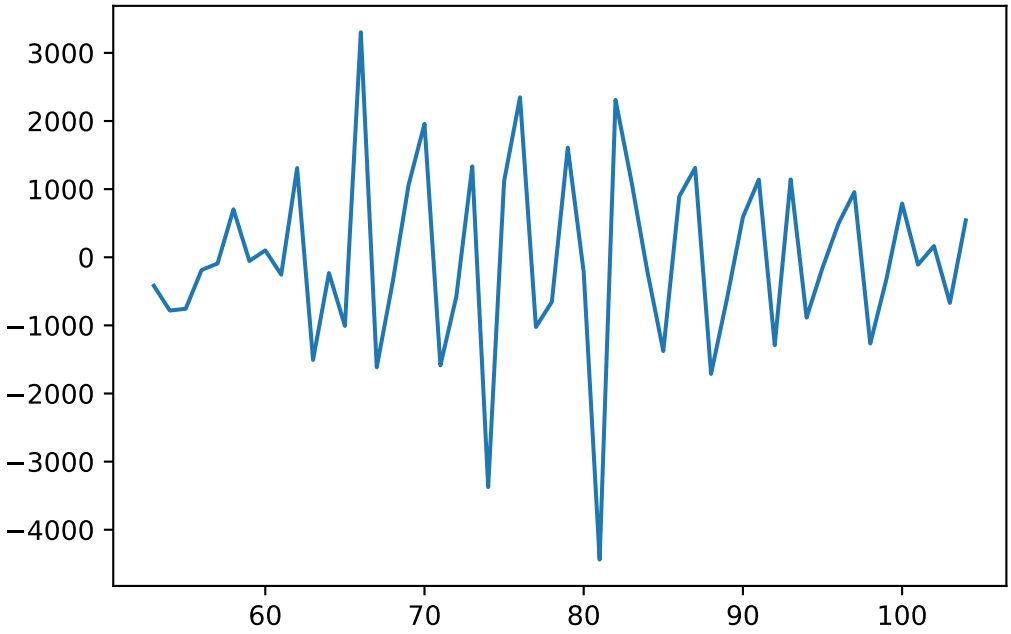
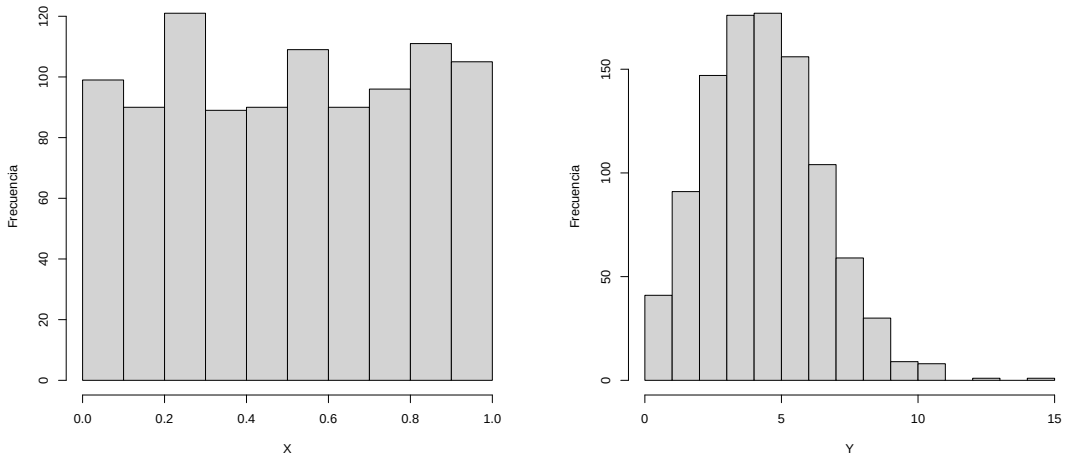


Figura 3: Convolución de las series de tiempo diferenciadas de casos registrados de enfermedades y concentraciones de partículas de 10 PM en el AMM durante 2017.



(a) X. (b) Y.

Figura 4: Histogramas de mil valores aleatorios para $X \sim \mathcal{U}(0, 1)$ y $Y \sim \mathcal{B}(100, 0.05)$.

```
cov(a * X + b, c * Y + d) == a * c * cov(X, Y)
# TRUE

var(X + Y) == var(X) + var(Y) + 2 * cov(X, Y)
# TRUE
```

Referencias

[1] Hyndman, R. J. y G. Athanasopoulos (2018), «Forecasting: Principles and Practice», `oTexts.com/fpp2`, [Accedido 13/nov/2020].

[2] Secretaría de Salud (2020), «Egresos Hospitalarios. Datos abiertos», URL `http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_egresoshosp_gobmx.html`.

[3] SISTEMA INTEGRAL DE MONITOREO AMBIENTAL [SIMA] NUEVO LEÓN (2020), «aire.nl.gob.mx | Home», `http://aire.nl.gob.mx/`, [Accedido 14/may/2020].