Distribuciones de probabilidad para segmentos de series temporales de conteos a tiempos discretos

Diego J. R. Sevilla †

† Facultad de Ciencias Exactas, Ingeniería y Agrimensura.
Universidad Nacional de Rosario.

E-mail: dsevilla@fceia.unr.edu.ar

Resumen

En este trabajo se presentan dos algoritmos para el cálculo de la probabilidad de segmentos de series temporales de conteos a tiempos discretos. Las distribuciones de probabilidad son calculadas mediante combinatoria, y pueden ser aplicadas a la detección de anomalías, especialmente para casos en que el número promedio de cuentas es bajo y las presuntas variaciones son débiles. Las distribuciones obtenidas son aplicadas a simulaciones, y los resultados comparados con los obtenidos de la función autocorrelación.

Introducción

En las últimas decadas fueron desarrollados gran cantidad de métodos estadísticos para el análisis de series temporales. Pero la detección de pequeñas variaciones aleatorias en las tasas de conteo son difíciles de detectar, y no existen recetas únicas para su tratamiento. En cualquier caso, el primer paso es detectar si elementos vecinos muestran correlación. Una importante herramienta es la función correlación [1], pero la información que brinda es limitada. En este trabajo se desarrollaron dos algoritmos para el cálculo de probabilidades de segmentos de datos, diseñados específicamente para procesos de conteo a tiempos discretos. El primero calcula la probabilidad de obtener n cuentas en m intervalos de tiempo, y el segundo la probabilidad de obtener un tiempo de espera n entre una cuenta y la m-ésima siguiente. Las distribuciones de probabilidad resultantes son las esperadas si la tasa de conteo fuese constante, es decir, si los elementos de la serie no mostraran correlación, y pueden ser utilizadas para determinar cuán estadísticamente probables son determinados segmentos de datos.

Método

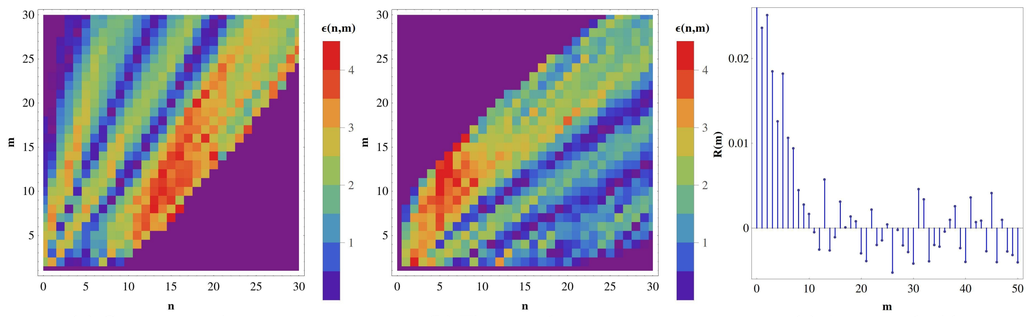
Los algoritmos fueron desarrollados en Wolfram *Mathematica*, y calculan mediante combinatoria la probabilidad $P^{(m)}(n)$ de obtener m elementos seguidos cuya suma sea n. Como distribución de probabilidad base para el cálculo combinatorio, se utilizan las estadísticas q_n de frecuencias de número de cuentas por intervalo de tiempo. Finalmente, se comparan las probabilidades $P^{(m)}(n)$ con las estadísticas $Q^{(m)}(n)$ de las sumas de m elementos seguidos mediante la siguiente función error:

$$\epsilon(n,m) = \frac{|Q^{(m)}(n) - P^{(m)}(n)|}{\sigma} \qquad \text{siendo} \qquad \sigma = \sqrt{\frac{P^{(m)}(n)\,(1 - P^{(m)}(n))}{N^{(m)}}}$$

Resultados

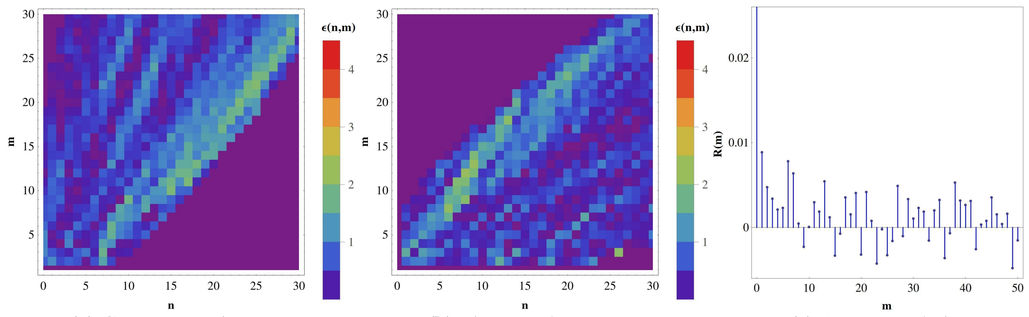
A continuación se presentan resultados de análisis de varios conjuntos de datos simulados. En las simulaciones se utilizaron dos distribuciones de probabilidad. Una con valor medio de cuentas `r0` para el estado base, y otra con valor medio de cuentas `rb` para las erupciones. Ambas distribuciones fueron generadas al azar. El número total de datos simulados fue `Ndata`, el número de erupciones `Nbursts` y la longitud de las erupciones `Lbursts`. Las erupciones se localizaron al azar.

A continuación se presentan resultados para erupciones de longitud 10.



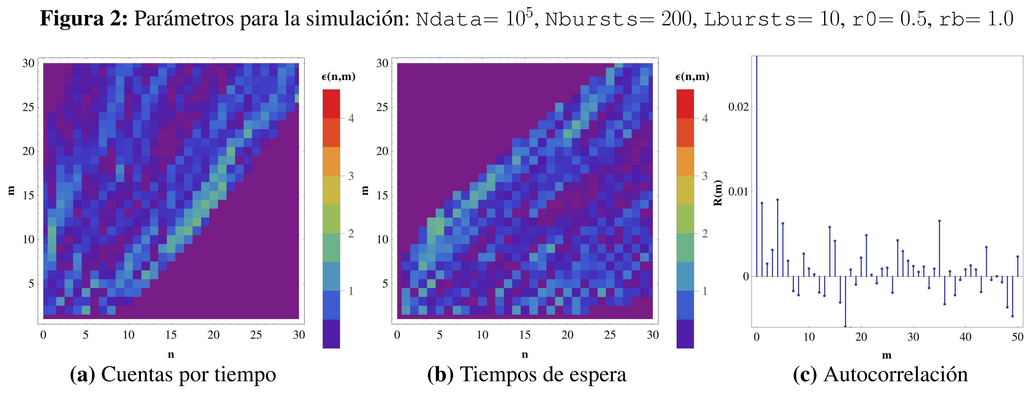
(a) Cuentas por tiempo **(b)** Tiempos de espera **(c)** Autocorrelación

Figura 1: Parámetros para la simulación: `Ndata`= 10^5, `Nbursts`= 500, `Lbursts`= 10, `r0`= 0.5, `rb`= 1.0



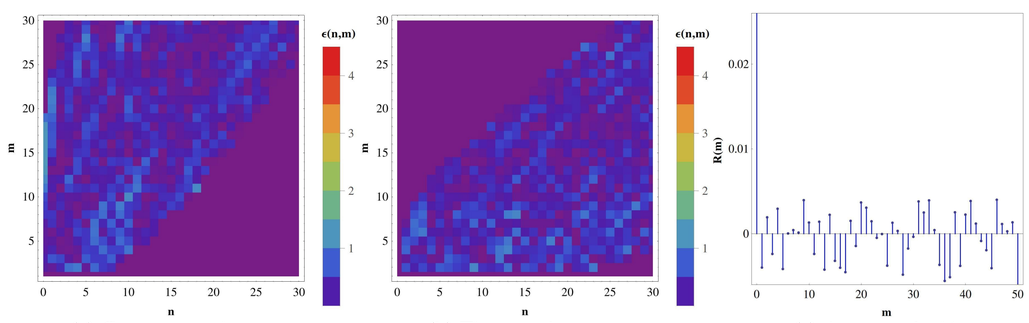
(a) Cuentas por tiempo **(b)** Tiempos de espera **(c)** Autoccorrelation

Figura 2: Parámetros para la simulación: `Ndata`= 10^5, `Nbursts`= 200, `Lbursts`= 10, `r0`= 0.5, `rb`= 1.0



(a) Cuentas por tiempo **(b)** Tiempos de espera **(c)** Autocorrelación

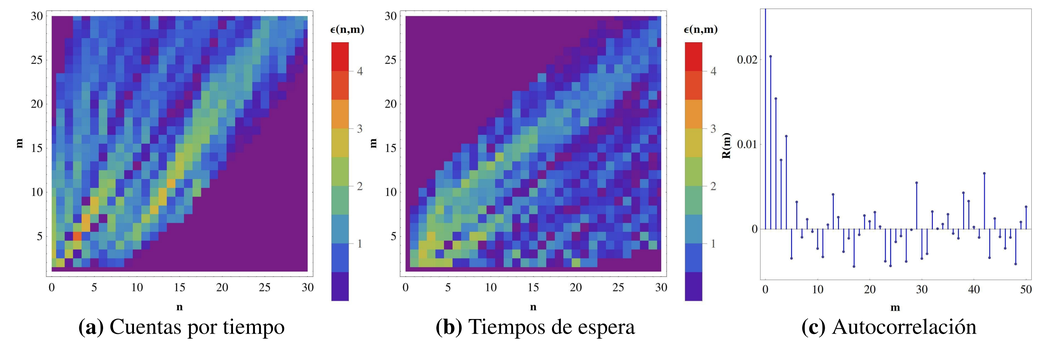
Figura 3: Parámetros para la simulación: `Ndata`= 10^5, `Nbursts`= 100, `Lbursts`= 10, `r0`= 0.5, `rb`= 1.0



(a) Cuentas por tiempo **(b)** Tiempos de espera **(c)** Autocorrelación

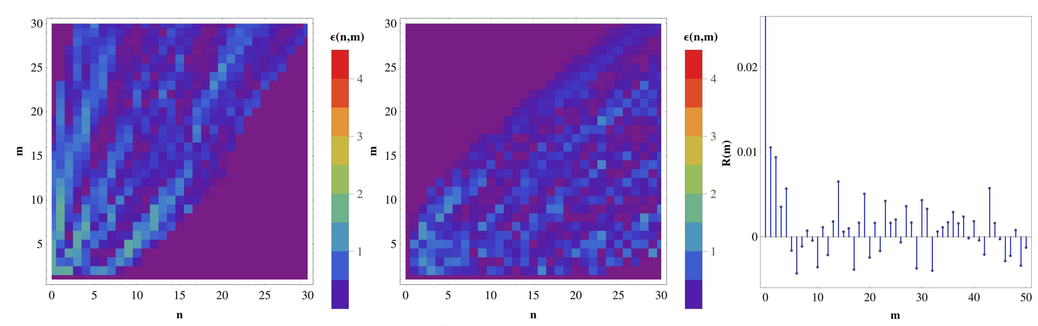
Figura 4: Parámetros para la simulación: `Ndata`= 10^5, `Nbursts`= 0, `Lbursts`= 10, `r0`= 0.5, `rb`= 1.0

Los siguientes resultados corresponden a simulaciones donde la longitud de las erupciones es 8, siendo los demás parámetros iguales.



(a) Cuentas por tiempo **(b)** Tiempos de espera **(c)** Autocorrelación

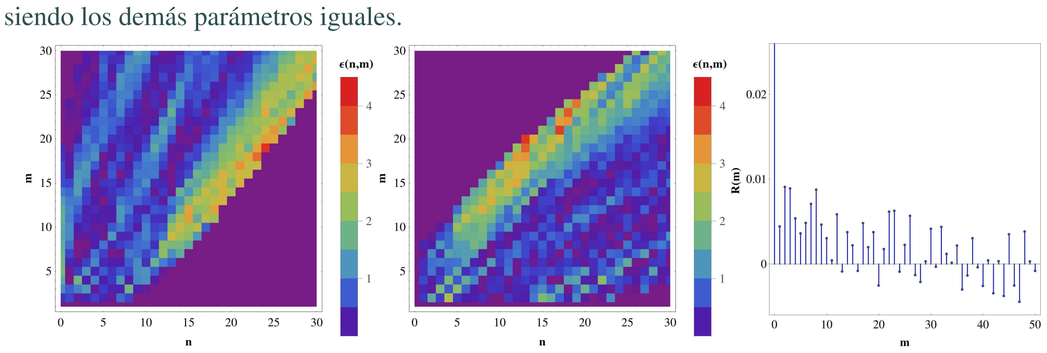
Figura 5: Parámetros para la simulación: `Ndata`= 10^5, `Nbursts`= 1000, `Lbursts`= 5, `r0`= 0.5, `rb`= 1.0



(a) Cuentas por tiempo **(b)** Tiempos de espera **(c)** Autoccorrelation

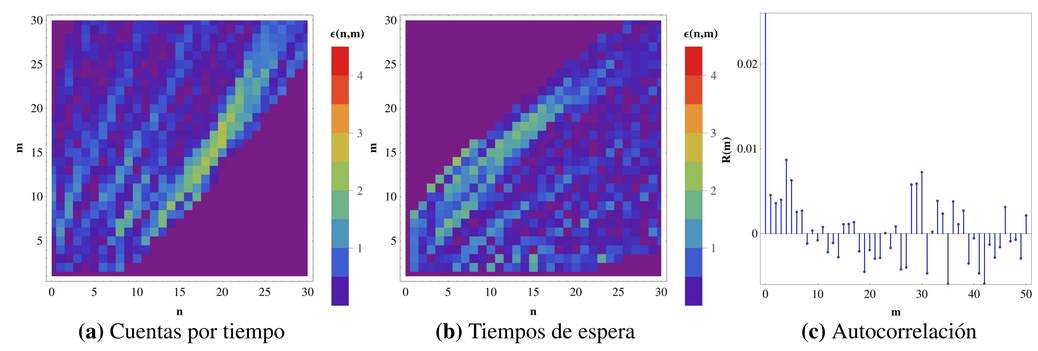
Figura 6: Parámetros para la simulación: `Ndata`= 10^5, `Nbursts`= 500, `Lbursts`= 5, `r0`= 0.5, `rb`= 1.0

Finalmente, se muestran resultados para simulaciones donde la longitud de las erupciones es 15, siendo los demás parámetros iguales.



(a) Cuentas por tiempo **(b)** Tiempos de espera **(c)** Autocorrelación

Figura 7: Parámetros para la simulación: `Ndata`= 10^5, `Nbursts`= 100, `Lbursts`= 15, `r0`= 0.5, `rb`= 1.0



(a) Cuentas por tiempo **(b)** Tiempos de espera **(c)** Autocorrelación

Figura 8: Parámetros para la simulación: `Ndata`= 10^5, `Nbursts`= 50, `Lbursts`= 15, `r0`= 0.5, `rb`= 1.0

Discusión

Las figuras 1 a 3 muestran el error antes definido en función de n y m para simulaciones con diferentes cantidades de erupciones de longitud 10. Se aprecia que a medida que el número de erupciones disminuye, el error decrece. En los resultados para la simulación con 100 erupciones (fig. 3), apenas llegan a apreciarse diferencias respecto a la simulación sin erupciones (fig. 4), es decir, puede considerarse que 100 prácticamente representa el número mínimo de erupciones que pueden ser detectadas en las condiciones dadas. Las figuras 5 y 6 muestran resultados para erupciones de longitud 8. Se aprecia que es necesario un mayor número de erupciones para que éstas puedan ser detectadas, siendo el número mínimo aproximadamente 500 para las condiciones dadas. Las figuras 7 y 8 muestran resultados para erupciones de longitud 15. Para este caso, el número mínimo de erupciones que pueden ser detectadas es aproximadamente 50.

Por otro lado, puede apreciarse que las regiones donde el error es mayor depende de la longitud de las erupciones. Este resultado podría ser utilizado para caracterizar las variaciones detectadas. Para ello es necesario extender la investigación, ya que se observa que la forma de las distribuciones de probabilidad utilizadas en las simulaciones también influye en la distribución del error.

Conclusiones

- A partir de las estadísticas de cuentas por intervalo de tiempo es posible calcular por combinatoria distribuciones de probabilidad para números totales de cuentas n en m intervalos de tiempo y de n tiempos de espera cada m cuentas.
- Las distribuciones de probabilidad pueden ser comparadas con las correspondientes estadísticas para detectar segmentos cuyo número supere a lo esperado.
- Los métodos puede ser útiles para detectar correlación entre elementos vecinos, como podría darse para erupciones aleatorias de baja intensidad.
- Ambos métodos muestran eficacias similares a la de la función correlación.

Referencias

[1] Patrick F Dunn. *Measurement and data analysis for engineering and science*. Crc Press, 2014.